INTERFACE

rsif.royalsocietypublishing.org

Review

Cite this article: Zoroa N, Lesigne E, Fernández-Sáez MJ, Zoroa P, Casas J. 2017 The coupon collector urn model with unequal probabilities in ecology and evolution. *J. R. Soc. Interface* **14**: 20160643. http://dx.doi.org/10.1098/rsif.2016.0643

Received: 12 August 2016
Accepted: 11 January 2017

Subject Category:
Life Sciences – Mathematics interface

Subject Areas:
biomathematics

Keywords:
coupon collector's problem, parasitoid, stochastic dominance, strong dominance, ecology

Author for correspondence:
N. Zoroa
e-mail: nzoroa@um.es

Electronic supplementary material is available online at https://dx.doi.org/10.6084/m9.figshare.c.3677185.

THE ROYAL SOCIETY PUBLISHING

The coupon collector urn model with unequal probabilities in ecology and evolution

N. Zoroa[1], E. Lesigne[2], M. J. Fernández-Sáez[1], P. Zoroa[1] and J. Casas[3]

[1]Departamento de Estadística e Investigación Operativa, Facultad de Matemáticas, Universidad de Murcia, 30071, Murcia, Spain
[2]Université de Tours, CNRS, LMPT UMR7350, Tours, France
[3]Université de Tours and Institut Universitaire de France Institut de Recherche en Biologie de l'Insecte, IRBI UMR CNRS 7261, Tours, France

NZ, 0000-0003-0937-8280

The sequential sampling of populations with unequal probabilities and with replacement in a closed population is a recurrent problem in ecology and evolution. Examples range from biodiversity sampling, epidemiology to the estimation of signal repertoire in animal communication. Many of these questions can be reformulated as urn problems, often as special cases of the coupon collector problem, most simply expressed as the number of coupons that must be collected to have a complete set. We aimed to apply the coupon collector model in a comprehensive manner to one example—hosts (balls) being searched (draws) and parasitized (ball colour change) by parasitic wasps—to evaluate the influence of differences in sampling probabilities between items on collection speed. Based on the model of a complete multinomial process over time, we define the distribution, distribution function, expectation and variance of the number of hosts parasitized after a given time, as well as the inverse problem, estimating the sampling effort. We develop the relationship between the risk distribution on the set of hosts and the speed of parasitization and propose a more elegant proof of the weak stochastic dominance among speeds of parasitization, using the concept of Schur convexity and the 'Robin Hood transfer' numerical operation. Numerical examples are provided and a conjecture about strong dominance—an ordering characteristic of random variables—is proposed. The speed at which new items are discovered is a function of the entire shape of the sampling probability distribution. The sole comparison of values of variances is not sufficient to compare speeds associated with different distributions, as generally assumed in ecological studies.

1. Introduction

The description of sequential sampling of a population of individuals for which the probability of being selected does not vary until a specific event, such as the collection of all or some types of individuals or a specific subgroup of the population, occurs is a common problem in ecology and evolution studies. In probability theory, such problems are often treated as urn problems, generally as the 'coupon collector problem' (CCP). The CCP is a mathematical model that belongs to the family of urn problems that can be formulated as follows: a company issues coupons of different types, each with a particular probability of being issued. The object of interest is the number of coupons that must be collected to obtain a full collection. This problem has been widely studied. The first findings concerned the classical problem in which all coupons are equally likely to be obtained [1]. Rapid advances have been made in this field [2–4], but they have gone largely unnoted by most scientists working in ecology and evolution. This is partly due to difficulties in making the correct analogies, partly due to a lack of worked examples and partly because each field devises its own vocabulary,

procedures and formalism. In ecological sciences, for example, a vibrant field of theoretical and applied ecological statistics developed in the 1950s from the repeated sampling of populations to estimate biodiversity richness [5,6]. This field could greatly benefit from the latest advances in the CCP [7,8] as we show through the working of a biological example. Related problems deal with the relative abundance of species from a community containing many species [9], or the sampling effort required to achieve a particular level of coverage [10]. Increases in the number of new hosts being infected or superinfected are a topic of great importance in population dynamics and epidemiology [11–14]. Several of the questions posed in capture–recapture studies relate to the coupon collector problem. Occupancy problems and related capture–recapture techniques are, indeed, defined as problems in which the probability of a given species occupying a given state at a given time must be determined (see the review [15] and the paper [16]). In ethological sciences, the estimation of a repertoire of signals in animal communication is considered as a form of the CCP, because vocal repertoire size is a key behavioural indicator of the complexity of the vocal communication system in birds and mammals [17]. In genetics and evolution, the CCP has been recognized as such only occasionally, despite these fields having generated some of the most elegant theorems and uses of other urn processes [18–20]. Indeed, the CCP has been used in the context of exhaustive haplotype sampling in phylogeography [21], determining the number of beneficial mutations as a function of sequence lines [22] and estimations of the size of the library required to target a particular percentage of the non-essential genome displaying a given property [23], for example.

Urn models have been much more widely used for modelling host–parasitoid systems than in other topics of ecology. We therefore used the biological context and formalism of parasitism by parasitic wasps, as the results obtained with this system can easily be extended to other ecological and evolutionary contexts. Parasitic wasps search for insects hosts, such as caterpillars, in which they lay a single, or multiple eggs. In solitary wasp species, only a single wasp develops fully in a given host. Parasitism can thus be formalized as a probabilistic dynamic process with hosts as 'balls' and parasitoids changing their 'colour' by parasitizing them. In work beginning more than a century ago [24,25], the pioneering population dynamicists assumed that hosts were found and attacked on successive occasions governed by exponential laws in continuous time. The number of draws was thus considered to be random and the number of eggs for a given host was assumed to follow a Poisson law [26]. If we assume that the number of draws is fixed, then the distribution of the number of eggs for a given host is binomial, but closely approximates a Poisson distribution in large host populations. The proportion of the population without eggs (the zero class) is of particular interest, because these hosts survive parasitism and produce offspring for the next generation. In field studies however, observed distributions are generally more aggregated than would be expected under the assumed Poisson distribution [27]. Aggregation is interpreted as the result of heterogeneity in the risk of being found, owing to differences in location, accessibility, appearance, colour, developmental stage or any other trait [28,29]. The risk distribution greatly influences the stability of the host–parasitoid system and has been widely studied [30–32].

All these works make strong assumptions about parasitoid searching and attacking behaviour and hence the egg distribution over the population of hosts, after a given time or a given number of draws (figure 1). However, the use of this distribution greatly decreases the amount of information available, as it collapses individual host histories. Parasitism is a multinomial process (figure 1), in which time corresponds to host draws. Its dynamics determines, for example, the percentage of hosts parasitized at the end of the season, the opportunity and time at which alternative pest control methods need to be deployed in supplement in biological control with parasitoids, and the time required to achieve a given degree of control by parasitic wasps. In this paper, we aimed to model parasitism as a multinomial urn process over time and we study the speed of parasitization (figure 1). We consider host encounters followed by oviposition without discrimination. The parasitism process described above can be considered as a CCP. In this case, there is a finite population of hosts differing in appearance, location, developmental stage or other factors. This heterogeneity results in different probabilities of hosts being found by parasitoids. These probabilities, p_h for host h, do not change over time. Our work therefore entails describing in depth the CCP, highlighting unnoticed analogies among previous works within the probability literature, and comparing the influence of the degree of heterogeneity among hosts on the speed of infection. We give a compact and hopefully more elegant proof than previously known of the following fact: the more the distribution p on the set of hosts is heterogeneous, the more the (random) number Y of parasitized hosts after a given number of draws is small; in other terms, there is a monotonic relationship between the majorization relation on the set of probability distributions p with the stochastic dominance on the set of random numbers Y.

This paper is structured as follows. In §2, we define a succession S_n, $n = 1,2,\ldots$, of N-dimensional random variables describing the state of the host population over time, in which time, n, is given by the number of attacks on the set of hosts. Each marginal distribution of S_n provides us information about a subset of hosts, including, in particular, the hth component representing the number of times that host h has been attacked by a parasitoid between times 1 and n. In §3, we define the random variables Y_n, $n = 1,2,\ldots$, representing the number of parasitized hosts after n draws. We also compute the distribution, the distribution function, the expectation and the variance of Y_n. We found no examples of calculations of this value in previous studies and therefore believe this aspect to be novel. We obtain the expected number of draws required for all hosts in a given subset to be parasitized and provide upper and lower bounds for this value in §4. In §5, we apply the results developed in previous sections to two particular risk distributions on the set of hosts. We first use the uniform distribution, and then a distribution corresponding to a host population with two different kinds of hosts. We calculate the most relevant values for each of these cases. In §6, we develop the relationship between the speed of parasitization and the risk distribution in the set of hosts. A narrower risk distribution is associated with faster parasitization. Thus, parasitization is fastest when the risk distribution is uniform. A conjecture on strong stochastic dominance is proposed in §7. In §8, we highlight the relationship between the speed of parasitization and the risk

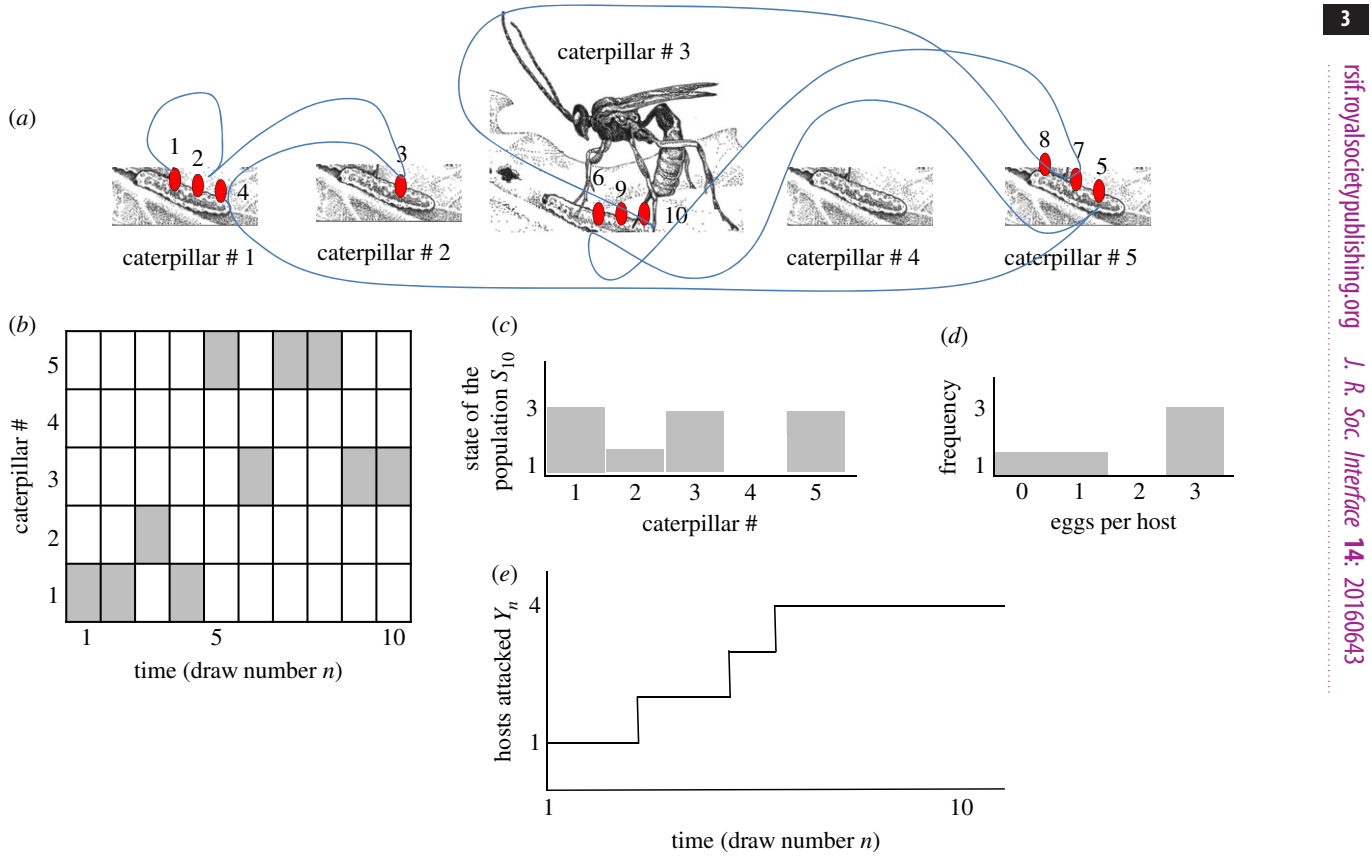

Figure 1. Attacks of caterpillar larvae hosts by parasitic wasps as an urn process in discrete time n. A wasp oviposits 10 eggs among five hosts ($n = 10$) (*a*). The outcome of the fundamental multinomial process (*b*) is summarized in the marginal distribution of the number of eggs per individual host at a given time $S_{10} = (3,1,3,0,3)$ (*c*), in the frequency distribution of eggs among hosts (*d*) and in the number of hosts attacked over time Y_n (*e*).

distribution with numerical examples and in §9 we apply our approach to a real host–parasitoid system. We have included the most mathematical aspects of the paper as electronic supplementary material.

2. Modelling parasitism as an urn process

We assume a finite population of hosts, constant for the entire duration of the experiment. The size of the parasitoid population is irrelevant, but we assume that the number of eggs that can be laid in the host population is not limiting. The situation is developed in successive stages or draws. At each stage, a parasitoid attacks a host, in which it lays an egg. The model is based on the fundamental assumption that successive draws are mutually independent. The hosts differ in appearance owing to intrinsic qualities, and these differences modify their probability of being attacked by a parasitoid. If the hosts are named 1, 2, 3, ..., N, then host h has a probability $p_h \geq 0$ ($\sum p_h = 1$) of being attacked by a parasitoid in a draw. This probability does not change during the process. We will say that $p_1, p_2, \ldots, p_N$ or $(p_1, p_2, \ldots, p_N)$ is the risk distribution for the set of hosts $H = \{1, 2, \ldots, N\}$.

The underlying probability space of our model is $(\Omega, \mathcal{S}, P)$, where the elements of Ω are all the possible histories of parasitism, that is $\Omega = H^{\mathbb{N}}$, equipped with its product σ-algebra $\mathcal{S}$ and the probability P given by Kolmogorov's theorem: if we therefore fix $i_1, i_2, i_3, \ldots, i_n$ in H, the probability of the event $\{\omega = i_1 i_2 i_3 \ldots i_n j_{n+1} j_{n+2} \ldots :$ for some j_k in $H, k > n\}$ is $p_{i_1} p_{i_2} \ldots p_{i_n}$. When necessary, the vector $(p_1, p_2, \ldots, p_N)$ will be denoted by a single letter p, and the probability P will be denoted by P_p.

We can describe this situation by defining a succession of random variables,

$$S_n = (S_{n1}, S_{n2}, \ldots, S_{nN}), \quad n = 1, 2, \ldots \tag{2.1}$$

where S_{nj} denotes the number of eggs in host j after n draws.

Variable S_n represents the state of the host population after n draws, that is, the distribution of eggs over the total population of hosts. If host i was visited r_i times between stages 1 and n, for $i = 1, 2, \ldots, N$, then S_n takes the value $(r_1, r_2, \ldots, r_N)$. This variable has a multinomial distribution with parameters n, $p_1, p_2, \ldots, p_N$, that is, for every integers $r_1, r_2, \ldots, r_N$, $0 \leq r_i \leq n$, $r_1 + r_2 + \cdots + r_N = n$,

$$P(S_{n1} = r_1, S_{n2} = r_2, \ldots, S_{nN} = r_N) = \frac{n!}{r_1! r_2! \ldots r_N!} p_1^{r_1} p_2^{r_2} \ldots p_N^{r_N}. \tag{2.2}$$

The marginal distribution of S_n, for $i_1, i_2, \ldots, i_h$ distinct elements of $\{1, 2, \ldots, N\}$ is given by

$$P(S_{ni_1} = r_{i_1}, S_{ni_2} = r_{i_2}, \ldots, S_{ni_h} = r_{i_h}) = \frac{n!}{r_{i_1}! r_{i_2}! \ldots r_{i_h}! (n - \sum r_{i_j})!} p_{i_1}^{r_{i_1}} p_{i_2}^{r_{i_2}} \ldots p_{i_h}^{r_{i_h}} q_{i_1 i_2 \ldots i_h}^{n - \sum r_{i_j}}, \tag{2.3}$$

with

$$0 \leq r_{i_j} \leq n, \quad j = 1, 2, \ldots, h, \quad r_{i_1} + r_{i_2} + \cdots + r_{i_h} \leq n,$$

$$q_{i_1 i_2 \ldots i_h} = 1 - \sum_{j=1}^{h} p_{i_j}.$$

This is the probability that, after n draws host i_1 has been visited r_{i_1} times by the parasitoids, host i_2 r_{i_2} times and host i_h r_{i_h} times, without considering the rest of the hosts.

In particular, the component S_{nh} of S_n has a binomial distribution with parameters n, p_h,

$$P(S_{nh}=r)=\frac{n!}{r!(n-r)!}p_{\mathrm{h}}^{r}q_{h}^{n-r},\quad r=0,1,2,\ldots,n \tag{2.4}$$

where

$$q_h=1-p_{\mathrm{h}}=\sum_{i\neq h}p_i.$$

This variable represents the state of host h after n draws. Thus, $P(S_{nh}=r)$ is the probability that host h has been attacked r times during the n draws.

The expected value and variance of this random variable are

$$E(S_{nh})=np_{\mathrm{h}},\quad \mathrm{Var}(S_{nh})=np_{\mathrm{h}}(1-p_{\mathrm{h}}).$$

Let $(e_1,e_2,\ldots,e_N)$ denote the canonical base of the space $\mathbb{R}^{\mathbb{N}}$. We emphasize that the process $(S_n)_{n\geq 1}$ is the random walk on Z_+^N with independent increments obeying the following law: $S_{n+1}-S_n=e_k$ with probability p_k. The statistical behaviour of this process is also very well known.

Note that, in this model, the sequence of random subsets of H, describing the set of parasitized hosts over time, is a Markov chain, and it is not difficult to give a precise description of its probability transitions. However, it is not straightforward to study this Markov chain directly.

3. Number of parasitized hosts after n draws

Let Y_n be the random variable representing the number of parasitized hosts after n draws, that is $Y_n=k$ if there are exactly k parasitized hosts after n draws. In this section, we give expressions for the probability mass function (3.3), distribution function (3.4), expectation (3.5) and variance (3.6) of this random variable. We show in annex A in electronic supplementary material how these expressions can be obtained.

From now on, for any integer $h>0$ and real x, we write

$$\left.\begin{aligned}&x^{(h)}=x(x-1)(x-2)\ldots(x-h+1),\quad x^{(0)}=1,\\&\binom{x}{h}=\frac{x^{(h)}}{h!}=\frac{x(x-1)\ldots(x-h+1)}{h!}\quad\text{and}\quad\binom{x}{0}=1.\end{aligned}\right\} \tag{3.1}$$

The distribution and the distribution function of Y_n have been obtained in previous studies, see [4]. Denoting $p_{j_1,j_2,\ldots,j_k}=p_{j_1}+p_{j_2}+\ldots+p_{j_k}$, the probability mass function is given by

$$\begin{aligned}P(Y_n=k)\;=&\sum_{\{j_1,j_2,\ldots,j_k\}\subset\{1,2,\ldots,N\}}p_{j_1j_2\ldots j_k}^{n}\\&-\binom{N-k+1}{N-k}\sum_{\{j_1,j_2,\ldots,j_{k-1}\}\subset\{1,2,\ldots,N\}}p_{j_1j_2\ldots j_{k-1}}^{n}\\&+\binom{N-k+2}{N-k}\sum_{\{j_1,j_2,\ldots,j_{k-2}\}\subset\{1,2,\ldots,N\}}p_{j_1j_2\ldots j_{k-2}}^{n}-\cdots\\&+(-1)^{k-1}\binom{N-1}{N-k}\sum_{\{j\}\subset\{1,2,\ldots,N\}}p_{j}^{n}.\end{aligned} \tag{3.2}$$

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

$$\left.\begin{aligned}&P(Y_n=k)=\sum_{A\subset H,|A|\leq k}(-1)^{k-|A|}\binom{N-|A|}{k-|A|}p_A^n,\\&\text{for } 0\leq k\leq\min\{N,n\},\end{aligned}\right\} \tag{3.3}$$

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

For the distribution function of Y_n, we obtain

$$\left.\begin{aligned}&P(Y_n\leq k)=\sum_{A\subset H,|A|\leq k}(-1)^{k-|A|}\binom{N-|A|-1}{k-|A|}p_A^n,\\&k=1,2,\ldots,N.\end{aligned}\right\} \tag{3.4}$$

A similar expression can be seen in [4].

It is well known that

$$E(Y_n)=N-m_{N-1}^{[n]}=N-\sum_{i=1}^{N}(1-p_i)^n, \tag{3.5}$$

but we have been unable to find any expression for $E(Y_n^2)$ and the variance of Y_n in previous studies. For these values, we obtain

$$\begin{aligned}E(Y_n^2)&=2m_{N-2}^{[n]}-(2N-1)m_{N-1}^{[n]}+N^2\\&=2\sum_{1\leq i<j\leq N}(1-p_i-p_j)^n-(2N-1)\sum_{i=1}^{N}(1-p_i)^n+N^2\end{aligned}$$

and

$$\begin{aligned}\mathrm{Var}(Y_n)=&\,2\sum_{1\leq i<j\leq N}(1-p_i-p_j)^n\\&+\sum_{i=1}^{N}(1-p_i)^n\left(1-\sum_{i=1}^{N}(1-p_i)^n\right)\end{aligned} \tag{3.6}$$

as can be seen in annex A in electronic supplementary material.

4. The number of draws required to reach a given level of parasitism

The expected number of draws required for the parasitization of k unparasitized hosts may be of considerable interest. For example, we might want to know the expected number of draws required for k of the hosts occupying a determined region, or with probabilities of parasitization greater (or less) than a given value, etc., to be parasitized. We define below a random variable representing the number of draws required for the event of interest to happen and we obtain its expectation. We also describe the relationship between the random variables defined here and the variables Y_n defined in §3.

Let us consider that, at a given stage of the process, there is a set $K\subset H$ of unparasitized hosts. This is our set of interest, and the remaining hosts $H-K$ are or are not parasitized. Let us use X to denote the number of hosts in the set $H-K$ attacked by the parasitoids before one of the hosts in K is attacked.

As this process involves the repeating of independent trials, the random variable X follows a geometric distribution with parameter $p=\sum_{i\in K}p_i$ (or a degenerate distribution if $K=H$). Thus,

$$E(X)=\frac{\sum_{i\in H-K}p_i}{\sum_{i\in K}p_i}. \tag{4.1}$$

Now, let k and N_1 be integers $1\leq k\leq N_1\leq N$. Let H_1 be a subset of the set of hosts, H, and $H_2=H-H_1$, $|H_1|=N_1$. We can assume that $H_1=\{1,2,\ldots N_1\}$ without loss of generality.

If we consider the hosts of set H_1 to be unparasitized, then we can define T_{k,N_1} as the random number of draws required to ensure that k hosts of set H_1 have been parasitized. Its expectation is the expected number of draws required for k hosts of set H_1 be parasitized. In this section, we include only the main ideas used to obtain this expectation, leaving the complete development for annex B in electronic

supplementary material. The case $H_1 = H$, and therefore $N_1 = N$, has been studied before and different expressions for $E(T_{k,N})$ have been obtained. We include these at the end of this section. In [2], an expression is proposed for the particular case in which $k = N_1 = N$.

Let $i_1, i_2, \ldots, i_k$ be distinct elements of H_1. Let $D_{i_1, i_2, \ldots, i_k}$ be the event defined by the fact that the first k hosts of set H_1 parasitized (i.e. attacked by a parasitoid for first time) are hosts $i_1, i_2, \ldots, i_k$, and are parasitized in the precise order i_1, $i_2, \ldots, i_k$. In other words, some of the hosts of set H_2 may be attacked first, followed by host i_1. Next, some hosts of $H_2 \cup \{i_1\}$ may be attacked, followed by host i_2, etc. Let $p = \sum_{i \in H_1} p_i$, $q = 1 - p = \sum_{i \in H_2} p_i$.

As proven in annex B in electronic supplementary material, we can write the equality

$$E(T_{k,N_1}) = \sum_{(i_1 i_2 \ldots i_k) \in \Pi_k} E(T_{k,N_1} \mid D_{i_1 i_2 \ldots i_k}) P(D_{i_1 i_2 \ldots i_k}), \tag{4.2}$$

where Π_k is the set of all k-permutations of $1, 2, \ldots, N_1$.

The probability of event $D_{i_1 i_2 \ldots i_k}$ and the conditional expectation $E(T_{k,N_1} | D_{i_1 i_2 \ldots i_k})$ are given by

$$P(D_{i_1 i_2 \ldots i_k}) = \frac{\prod_{j=1}^{k} p_{i_j}}{p(p - p_{i_1})(p - p_{i_1} - p_{i_2}) \ldots (p - \sum_{j=1}^{k-1} p_{i_j})} \tag{4.3}$$

and

$$\begin{aligned} E(T_{k,N_1} \mid D_{i_1 i_2 \ldots i_k}) &= \left(\sum_{h=1}^{k} \frac{q + \sum_{j=1}^{h-1} p_{i_j}}{p - \sum_{j=1}^{h-1} p_{i_j}} \right) + k = \sum_{h=1}^{k} \left(\frac{q + \sum_{j=1}^{h-1} p_{i_j}}{p - \sum_{j=1}^{h-1} p_{i_j}} + 1 \right) \\ &= \frac{1}{p} + \frac{1}{p - p_{i_1}} + \frac{1}{p - p_{i_1} - p_{i_2}} + \cdots + \frac{1}{p - \sum_{j=1}^{k-1} p_{i_j}} \\ &= \frac{1}{1-q} + \frac{1}{1 - q - p_{i_1}} + \cdots + \frac{1}{1 - q - \sum_{j=1}^{k-1} p_{i_j}}. \end{aligned} \tag{4.4}$$

Bearing in mind (4.2)–(4.4), we can state the following.

Proposition 4.1. *The expected value of T_{k,N_1} is*

$$E(T_{k,N_1}) = \sum_{(i_1 i_2 \ldots i_k) \in \Pi_k} \left(\frac{1}{p} + \frac{1}{p - p_{i_1}} + \frac{1}{p - p_{i_1} - p_{i_2}} + \cdots + \frac{1}{p - \sum_{j=1}^{k-1} p_{i_j}} \right) \frac{\prod_{j=1}^{k} p_{i_j}}{p(p - p_{i_1})(p - \sum_{j=1}^{2} p_{i_j}) \ldots (p - \sum_{j=1}^{k-1} p_{i_j})}, \tag{4.5}$$

where Π_k is the set of all k-permutations of set $\{1, 2, \ldots, N_1\}$, i.e. the arrangements of length k of different elements of $\{1, 2, \ldots, N_1\}$.

Thus, $E(T_{k,N_1})$ given by (4.5) is the expected number of draws required for k hosts of a set of unparasitized hosts $H_1 \subset H$ with cardinality N_1, to be parasitized. This value is generally difficult to obtain, because the number of terms required for its computation is the number of k-permutations of $1, 2, \ldots, N_1$, that is $N_1^{(k)} = N_1(N_1 - 1) \ldots (N_1 - k + 1)$. This value is huge when N_1 and k are large. It is therefore important to obtain upper and lower bounds for this value.

Proposition 4.2. *Let k be given and $p_1, p_2, \ldots, p_{N_1}$ be real numbers satisfying $p_1 \geq p_2 \geq \cdots \geq p_{N_1}$. Then, the maximum of $E(T_{k,N_1} \mid D_{i_1 i_2 \ldots i_k})$ defined by (4.4) over all possible choices of the k-subsets $\{i_1, i_2, \ldots, i_k\}$ of H_1 is*

$$E(T_{k,N_1} | D_{1,2,\ldots,k}) = \frac{1}{\sum_{i=1}^{N_1} p_i} + \frac{1}{\sum_{i=2}^{N_1} p_i} + \cdots + \frac{1}{\sum_{i=k}^{N_1} p_i}, \tag{4.6}$$

and the minimum is

$$E(T_{k,N_1} | D_{N_1, N_1 - 1, \ldots, N_1 - k + 1}) = \frac{1}{\sum_{i=1}^{N_1} p_i} + \frac{1}{\sum_{i=1}^{N_1 - 1} p_i} + \cdots + \frac{1}{\sum_{i=1}^{N_1 - k + 1} p_i}. \tag{4.7}$$

Proof. From hypothesis $p_1 \geq p_2 \geq \cdots \geq p_{N_1}$, it follows directly that

$$\sum_{i=h}^{N_1} p_i \leq \sum_{j=h}^{N_1} p_{i_j} \leq \sum_{i=1}^{N_1 - h + 1} p_i, \quad h = 1, 2 \ldots, N_1, \tag{4.8}$$

then

$$\begin{aligned} E(T_{k,N_1} | D_{1,2,\ldots,k}) &= \frac{1}{p} + \frac{1}{p - p_1} + \frac{1}{p - \sum_{i=1}^{2} p_i} + \cdots + \frac{1}{p - \sum_{i=1}^{k-1} p_i} \\ &\geq \frac{1}{p} + \frac{1}{p - p_{i_1}} + \frac{1}{p - p_{i_1} - p_{i_2}} + \cdots + \frac{1}{p - \sum_{j=1}^{k-1} p_{i_j}} \\ &\geq \frac{1}{p} + \frac{1}{p - p_{N_1}} + \frac{1}{p - \sum_{i=N_1 - 1}^{N_1} p_i} + \cdots + \frac{1}{p - \sum_{i=N_1 - k + 2}^{N_1} p_i} \\ &= E(T_{k,N_1} | D_{N_1, N_1 - 1, \ldots, N_1 - k + 1}) \end{aligned}$$

and the proof is complete. ■

Proposition 4.3. *Let $p_1, p_2, \ldots, p_{N_1}$ be real numbers satisfying $0 \leq p_i \leq 1$, for $i = 1, 2, \ldots, N_1$ and $p_1 \geq p_2 \geq \cdots \geq p_{N_1}$. It is then true that*

$$E(T_{k,N_1} \mid D_{1,2,\ldots,k}) \geq E(T_{k,N_1}) \geq E(T_{k,N_1} \mid D_{N_1, N_1 - 1, \ldots N_1 - k + 1}).$$

In other words, $E(T_{k,N_1} | D_{1,2,\ldots,k})$ and $E(T_{k,N_1} \mid D_{N_1, N_1 - 1, \ldots N_1 - k + 1})$ are upper and lower bounds, respectively, for the expected number of draws required for k hosts of the set H_1 to be parasitized.

Furthermore, the mode of the distribution on the events $D_{i_1 i_2 \ldots i_k}$, $(i_1, i_2, \ldots, i_k) \in \Pi_k$, is $D_{1,2,\ldots,k}$, i.e. the order of parasitism of k hosts mostly likely to occur is $1, 2, \ldots, k$.

Proof. The first part of this proposition is a straightforward consequence of the proposition 4.2.

The second part comes directly from the fact that

$$P(D_{1,2,\ldots,k}) \geq P(D_{i_1 i_2 \ldots i_k}) \quad \text{for } (i_1 i_2 \ldots i_k) \in \Pi_k,$$

which follows from (4.3) and (4.8). ■

Propositions 4.2 and 4.3 prove that, if $p_1 \geq p_2 \geq \cdots \geq p_{N_1}$, then the most likely order of parasitization of k hosts in H_1 is

the preferential order 1, 2, ..., k. Moreover, the quickest scenario (in terms of expectation) for the parasitization of k hosts of H_1 is the sequence extending from the least likely host, N_1, to the most likely host, $N_1 - k + 1$, in the correct order. The slowest scenario (in terms of expectation) for the parasitization of k hosts of H_1 extends from the most likely, 1, to the least likely host, k, in the correct order.

These results can be intuitively explained as follows. Let us suppose that host 1 is parasitized in the first place. The probability of a new host of the set $H_1 - \{1\}$ being parasitized is then $q - p_1$. This value is less than any other value $q - p_j$ with $j \neq 1$. It is therefore more difficult for a host of the set $H_1 - \{1\}$ to be parasitized than for a host of the set $H_1 - \{j\}, j \neq 1$, to be parasitized. The repeated application of this reasoning explains the first inequality of the proposition. The second inequality can be explained in a similar manner.

For simplicity, we denote $T_{k,N}$ by T_k in the particular case in which $N_1 = N$. Now, recalling the definitions of these random variables and the random variables Y_n, and bearing in mind the equivalence between 'The number of parasitized hosts after n draws is less than or equal to $k-1$ in the history of the sequence of parasitism', and 'To parasitize k hosts, more than n draws are necessary', we can write the equality

$$P(Y_n \leq k-1) = P(T_k > n),$$

and from this

$$P(Y_n \leq k-1) = 1 - P(T_k \leq n)$$

and

$$P(T_k = n) = P(Y_{n-1} \leq k-1) - P(Y_n \leq k-1).$$

From (3.3), (3.4) and above equalities, we see that

$$\left.\begin{aligned} P(T_k > n) &= \sum_{A \subset H, |A| \leq k-1} (-1)^{k-1-|A|} \binom{N-|A|-1}{k-1-|A|} p_A^n \\ \text{and } P(T_k \leq n) &= 1 - \sum_{A \subset H, |A| \leq k-1} (-1)^{k-1-|A|} \binom{N-|A|-1}{k-1-|A|} p_A^n \end{aligned}\right\} \tag{4.9}$$

and

$$\begin{aligned} P(T_k = n) &= \sum_{A \subset H, |A| \leq k-1} (-1)^{k-1-|A|} \binom{N-|A|-1}{k-1-|A|} p_A^{n-1} \\ &\quad - \sum_{A \subset H, |A| \leq k-1} (-1)^{k-1-|A|} \binom{N-|A|-1}{k-1-|A|} p_A^n \\ &= \sum_{A \subset H, |A| \leq k-1} (-1)^{k-1-|A|} \binom{N-|A|-1}{k-1-|A|} p_A^{n-1}(1-p_A). \end{aligned}$$

$E(T_k) = E(T_{k,N})$ is the expected number of draws required for k hosts to be parasitized. Different expressions have been described for this expectation [2,3]. From (4.9), it follows immediately that

$$\begin{aligned} E(T_k) &= \sum_{n=0}^{\infty} P(T_k > n) \\ &= \sum_{n=0}^{\infty} \left(\sum_{A \subset H, |A| \leq k-1} (-1)^{k-1-|A|} \binom{N-|A|-1}{k-1-|A|} p_A^n \right) \\ &= \sum_{A \subset H, |A| \leq k-1} (-1)^{k-1-|A|} \binom{N-|A|-1}{k-1-|A|} \frac{1}{1-p_A}. \end{aligned}$$

This expression was obtained in [3]. In [2], the following expression was obtained:

$$E(T_k) = \sum_{r=0}^{k-1} \|u^r\| \int_{t \geq 0} \prod_{i=1}^{N} (1 + u(e^{p_i t} - 1)) e^{-t} \, dt,$$

where $\|x^r\| f(x)$ is the coefficient of x^r in the power series development of $f(x)$.

If $k = N_1 = N$, then $E(T_N) = E(T_{N,N})$ is the expected number of draws required to obtain complete parasitism. From (4.5)

$$E(T_N) = \sum_{(i_1 i_2 \ldots i_N) \in \Pi_N} \left(\sum_{r=0}^{N-1} \frac{1}{1 - \sum_{j=1}^{r} p_{i_j}} \right) \frac{\prod_{i=1}^{N} p_i}{\prod_{k=1}^{N} \sum_{j=k}^{N} p_{i_j}}, \tag{4.10}$$

where Π_N is the group of permutations of $\{1, 2, \ldots N\}$. This expression for $E(T_N)$ is proposed in [2]. The authors provide no proof for this formula, and we have found no proof elsewhere.

5. Applications to various risk distributions

In this section, we consider two different risk distributions on the set of hosts and compute the most relevant values for every distribution.

5.1. The uniform distribution

The situation in which the risk is distributed uniformly, i.e. all the hosts have the same probability of being parasitized, with

$$p_1 = p_2 = \cdots = p_N = \frac{1}{N} \tag{5.1}$$

has been widely studied. In this case, the expectation and variance of the random variable Y_n representing the number of parasitized hosts after n draws are

$$\begin{aligned} E(Y_n) &= N - \frac{(N-1)^n}{N^{n-1}}, \\ \mathrm{Var}(Y_n) &= \frac{(N-1)(N-2)^n}{N^{n-1}} + \frac{(N-1)^n (N^{n-1} - (N-1)^n)}{N^{2n-2}}, \end{aligned}$$

and the expected number of draws for k new hosts to be parasitized (4.5) is

$$E(T_{k,N_1}) = N\left(\frac{1}{N_1} + \frac{1}{N_1 - 1} + \cdots + \frac{1}{N_1 - k + 1} \right),$$

which, in the case in which $k = N$, can be written as the following well-known formula:

$$E(T_N) = N\left(1 + \frac{1}{2} + \frac{1}{3} \ldots + \frac{1}{N}\right).$$

It is clear that in this case the upper and lower bounds for $E(T_{k,N_1})$ obtained in proposition 4.3 are both equal to $E(T_{k,N_1})$, and all the probabilities $P(D_{i_1 i_2 \ldots i_k})$ are equal to $1/N_1^{(k)}$.

5.2. Two kinds of hosts

The two types of host situations are an idealization of the following cases. Hosts which are dead, either because they were previously parasitized or because they produced artefacts such as mines and galls, remain in the ecosystem for much longer than the existence of the host. They can make up to 90% of the host population. They can be still attractive to parasitoids long after the host death. Parasitoids will not lay eggs in them, but they will be checked carefully, implying a waste of time of up to 20% [33,34]. In such cases, it is possible to envision two categories, living and dead hosts,

while being interested in the rate of parasitism of the living ones only.

Let us now consider the situation in which there are two kinds of hosts and, therefore, two different probabilities of being detected by a parasitoid.

In a population of N hosts, each of the hosts $1,2,\ldots,m$ has a probability α of being parasitized, and each host $m+1,m+2,\ldots,N$ has a probability β of being parasitized, such that

$$\left.\begin{array}{rc} & p_1=p_2=\cdots=p_m=\alpha \\ \text{and} & p_{m+1}=p_{m+2}=\cdots=p_N=\beta. \end{array}\right\} \tag{5.2}$$

The probability of host 1 being visited r_1 times, host 2 r_2 times, etc., for $r_1+r_2+\cdots+r_N=n$, given by (2.2) is in this case

$$P(S_{n1}=r_1,S_{n2}=r_2,\ldots,S_{nN}=r_N)=\frac{n!}{r_1!r_2!\ldots r_N!}\alpha^{\sum_{i\le m} r_i}\beta^{\sum_{i\le m} r_i}$$
$$0\le r_1\le n,0\le r_2\le n,\ldots,0\le r_N\le n,\ r_1+r_2+\cdots+r_N=n.$$

The probability that, after n draws host i_1 had been chosen r_{i_1} times by the parasitoids, host i_2 r_{i_2} times and host i_h r_{i_h} times, without taking the other hosts into account, is given by (2.3). It is equal to

$$P(S_{ni_1}=r_{i_1},S_{ni_2}=r_{i_2},\ldots,S_{ni_h}=r_{i_h})$$
$$=\frac{n!}{r_{i_1}!r_{i_2}!\ldots r_{i_h}!(n-\sum r_{i_j})!}\alpha^{\sum_{i_j\le m} r_{i_j}}\beta^{\sum_{i_j>m} r_{i_j}}\left(1-\sum_{i_j\le m}\alpha-\sum_{i_j>m} r_{i_j}\beta\right)^{n-\sum r_{i_j}}.$$

We now calculate the expected number of parasitized hosts after n draws with this risk distribution, using the results obtained in §2.

Let Y_n be the random variable representing the number of parasitized hosts after n draws. From (3.3), it follows that

$$P(Y_n=k)=\sum_{j=1}^{k}(-1)^{k-j}\binom{N-j}{k-j}\sum_{i=0}^{j}\binom{m}{i}\binom{N-m}{j-i}$$
$$(i\alpha-(j-i)\beta)^n$$

and the expected value of Y_n, (3.5), is equal to

$$E(Y_n)=N-m(1-\alpha)^n-(N-m)(1-\beta)^n.$$

To compute the expected number of draws for k hosts of a set $H_1\subset H$ of unparasitized hosts to be parasitized, we name the hosts of the set H_1, hosts $1,2,\ldots,N_1$. Without any loss of generality, we can assume $p_1=p_2=\cdots=p_{m_1}=\alpha$ and $p_{m_1+1}=p_{m_1+2}=\cdots=p_{N_1}=\beta$. Let Π_k be the set of all k-permutations of the integers $1,2,\ldots,N_1$. For every $I=(i_1,i_2,\ldots,i_k)\in\Pi_k$, let $A_I\subset\{1,2,\ldots,k\}$ be the set defined by $j\in A_I$ if $i_j\le m_1$. It is clear that the probability $P(D_{i_1,i_2,\ldots,i_k})=P(D_I)$ given by (4.3) is, in this case,

$$P(D_I)=\frac{\alpha^{|A_I|}\beta^{k-|A_I|}}{p(p-\gamma_1)(p-\sum_{j=1}^{2}\gamma_j)\ldots(p-\sum_{j=1}^{k-1}\gamma_j)},$$

where

$$\gamma_j=\begin{cases}\alpha & \text{if } j\in A_I\\ \beta & \text{if } j\notin A_I.\end{cases} \tag{5.3}$$

Then, if $A_I=A_{I'}$ for $I\in\Pi_k$ and $I'\in\Pi_k$, it follows directly that

$$P(D_I)=P(D_{I'}).$$

We can therefore define an equivalence relation on Π_k as follows: I is related to I' if $A_I=A_{I'}$. We denote by $\bar{I}$ the equivalence class of I, and by $\bar{\Pi}_k$ the set whose elements are the equivalence classes of the elements of Π_k, that is

$$\bar{\Pi}_k=\{\bar{I}:I\in\Pi_k\}.$$

There are as many equivalence classes as subsets of $\{1,2,\ldots,k\}$ with cardinalities greater than or equal to $\max\{0,k-n_1\}$, where $n_1=N_1-m_1$, and less than or equal to $\min\{k,m_1\}$, and the cardinalities of these equivalence classes are

$$|\bar{I}|=m_1^{(h)}n_1^{(k-h)}\quad\text{if } |A_I|=h.$$

Given the above considerations, it is clear that $E(T_{k,N_1})$ can be written in this case as

$$\begin{aligned}E(T_{k,N_1})&=\sum_{I\in\Pi_k}\left(\frac{1}{p}+\frac{1}{(p-\gamma_1)}+\cdots+\frac{1}{\left(p-\sum_{j=1}^{k-1}\gamma_j\right)}\right)\\&\quad\frac{\alpha^{|A_I|}\beta^{k-|A_I|}}{p(p-\gamma_1)\left(p-\sum_{j=1}^{2}\gamma_j\right)\ldots\left(p-\sum_{j=1}^{k-1}\gamma_j\right)}\\&=\sum_{\bar{I}\in\bar{\Pi}_k}\sum_{I\in\bar{I}}\left(\frac{1}{p}+\frac{1}{(p-\gamma_1)}+\cdots+\frac{1}{\left(p-\sum_{j=1}^{k-1}\gamma_j\right)}\right)\\&\quad\frac{\alpha^{|A_I|}\beta^{k-|A_I|}}{p(p-\gamma_1)\left(p-\sum_{j=1}^{2}\gamma_j\right)\ldots\left(p-\sum_{j=1}^{k-1}\gamma_j\right)}\\&=\sum_{\bar{I}\in\bar{\Pi}_k}m_1^{(|A_I|)}n_1^{(k-|A_I|)}\left(\frac{1}{p}+\frac{1}{(p-\gamma_1)}+\cdots+\frac{1}{\left(p-\sum_{j=1}^{k-1}\gamma_j\right)}\right)\\&\quad\frac{\alpha^{|A_I|}\beta^{k-|A_I|}}{p(p-\gamma_1)\left(p-\sum_{j=1}^{2}\gamma_j\right)\ldots\left(p-\sum_{j=1}^{k-1}\gamma_j\right)}\\&=\sum_{h=\max\{0,k-n_1\}}^{\min\{k,m_1\}}\sum_{|A_I|=h}m_1^{(h)}n_1^{(k-h)}\\&\quad\left(\frac{1}{p}+\frac{1}{(p-\gamma_1)}+\cdots+\frac{1}{(p-\sum_{j=1}^{k-1}\gamma_j)}\right)\\&\quad\frac{\alpha^h\beta^{k-h}}{p(p-\gamma_1)\left(p-\sum_{j=1}^{2}\gamma_j\right)\ldots\left(p-\sum_{j=1}^{k-1}\gamma_j\right)},\end{aligned} \tag{5.4}$$

where γ_j is defined by (5.3).

Let us suppose that

$$\alpha>\beta.$$

To obtain an upper bound for $E(T_{k,N_1})$, we distinguish two cases, $k\le m_1$ and $k>m_1$. If $k\le m_1$ then

$$E(T_{k,N_1}\mid D_{1,2,\ldots,k})=\frac{1}{m_1\alpha+n_1\beta}+\frac{1}{(m_1-1)\alpha+n_1\beta}+\cdots$$
$$+\frac{1}{(m_1-k+1)\alpha+n_1\beta}.$$

If $k > m_1$, this upper bound is

$$E(T_{k,N_1}|D_{1,2,\ldots,k}) = \frac{1}{m_1\alpha + n_1\beta} + \frac{1}{(m_1-1)\alpha + n_1\beta} + \cdots + \frac{1}{n_1\beta} + \frac{1}{(n_1-1)\beta} + \cdots + \frac{1}{(n_1+m_1-k+1)\beta}.$$

Similarly, to obtain a lower bound for $E(T_{k,N_1})$ we distinguish the cases $k \leq n_1$ and $k > n_1$. If $k \leq n_1$, this lower bound is

$$E(T_{k,N_1}|D_{N_1,N_1-1,\ldots,N_1-k+1}) = \frac{1}{m_1\alpha + n_1\beta} + \frac{1}{m_1\alpha + (n_1-1)\beta} + \cdots + \frac{1}{m_1\alpha + (n_1-k+1)\beta},$$

and if $k > n_1$, a lower bound for $E(T_{k,N_1})$ is

$$E(T_{k,N_1}|D_{N_1,N_1-1,\ldots,N_1-k+1}) = \frac{1}{m_1\alpha + n_1\beta} + \frac{1}{m_1\alpha + (n_1-1)\beta} + \cdots + \frac{1}{m_1\alpha} + \frac{1}{(m_1-1)\alpha} + \frac{1}{(n_1+m_1-k+1)\alpha}.$$

The maximum of the values $P(D_{i_1,i_2,\ldots,i_k})$ is

$$P(D_{1,2,\ldots,k}) = \begin{cases} \dfrac{\alpha^k}{\prod_{h=0}^{k-1}((m_1-h)\alpha + n_1\beta)}, & \text{if } k \leq m_1 \\ \dfrac{\alpha^{m_1}\beta^{k-m_1}}{\prod_{h=0}^{m_1}((m_1-h)\alpha + n_1\beta)\prod_{l=1}^{k-m_1-1}(n_1-l)\beta}, & \text{if } k > m_1. \end{cases}$$

6. Relationship between the risk distribution and the speed of parasitization

In the preceding sections, we studied the process of parasitization for a given risk distribution in the set of hosts. In this section, we compare this process for different risk distributions. We show how parasitization speed depends on the risk distribution, and its scatter in particular. We use the concept of 'majorization' to formalize the idea that risk distributions have different degrees of spread. This notion dates from the start of the twentieth century. A comprehensive review of the theory can be found in [35].

Less spread distributions are associated with faster parasitization. In other words, the more spread out the risk distribution, the larger the number of draws required for a given number of hosts to be parasitized. Thus, the distribution function for the first-time parasitization of a given number of hosts, viewed as a function of the vector p, is Schur convex (see the definition at the end of this section). The mathematical community studying the CCP seems to be largely unaware of it, but this result is not new and can be found in [36]. This result constitutes the first part of theorem 6.3. We give a proof which is more concise and clearer than a previous proposal. It is contained in annex C in supplementary electronic material.

Moreover, our method provides a precise result for strict Schur convexity. This refinement constitutes the second part of theorem 6.3. We make use in our proof of the relationship between the concept of majorization and the numerical operation known as 'Robin Hood transfer', described below.

In this section, we work with different risk distributions, requiring further notation and definitions. Given a risk distribution $p = (p_1, p_2, \ldots, p_N)$, we denote by P_p the probability distribution induced by p on the σ-field over the space of the all the possible incidences of parasitization.

Given $(p_1, p_2, \ldots, p_N)$ in $\mathbb{R}^N$, we denote by $(p_{\bar{1}}, p_{\bar{2}}, \ldots, p_{\bar{N}})$ the N-uple obtained by permutation of p_i such that $p_{\bar{1}} \geq p_{\bar{2}} \geq \cdots \geq p_{\bar{N}}$.

The following definitions are given in [35].

Definition 6.1. Let $p_1, p_2, \ldots, p_N$, $q_1, q_2, \ldots, q_N$, be real numbers. We say that $p = (p_1, p_2, \ldots, p_N)$ is majorized by $q = (q_1, q_2, \ldots, q_N)$, and we write $p \prec q$, if

$$\sum_{i=1}^{k} p_{\bar{i}} \leq \sum_{i=1}^{k} q_{\bar{i}} \quad \text{for } i = 1, 2, \ldots, N-1$$

and

$$\sum_{i=1}^{N} p_{\bar{i}} = \sum_{i=1}^{N} q_{\bar{i}}.$$

It is clear that when we apply this definition to the comparison of two risk distributions, the last equality is trivially satisfied.

Let $q = (q_1, q_2, \ldots, q_N) \in \mathbb{R}^N$. If $q_h < q_k$, then we can transfer an amount Δ, $0 < \Delta < q_k - q_h$ from q_k to q_h to obtain the following new risk distribution $q' = (q'_1, q'_2, \ldots, q'_N)$, where $q'_h = q_h + \Delta$, $q'_k = q_k - \Delta$ and $q'_i = q_i$ for $i \neq h, k$. Then, q' is less spread out than the initial distribution, that is, $q' \prec q$. Such operations involving the shifting of some 'income' from one individual to a poorer individual are described, somewhat poetically, as Robin Hood transfers [37]. If we define $\alpha = 1 - \Delta/(q_k - q_h)$, then we can write $q'_h = q_h + \Delta = \alpha q_h + (1-\alpha)q_k$ and $q'_k = q_k - \Delta = \alpha q_k + (1-\alpha)q_h$.

Proposition 6.2. *The following conditions are equivalent:*

(a) $p \prec q$ *and*
(b) *p can be derived from q by successive applications of a finite number of Robin Hood transfers.*

It is not difficult to prove this equivalence. It was proved for the first time, to the best of our knowledge, in [38] for vectors of non-negative integer components.

Let $p = (p_1, p_2, \ldots, p_N)$ denote a probability distribution p over the set H. Suppose that p is not uniform. We can assume $p_1 < p_2$ without loss of generality. Let $0 < h \leq (p_2 - p_1)/2$, $\alpha = 1 - h/(p_2 - p_1)$. We then define a new risk distribution p' by applying a Robin Hood transfer as follows:

$$p' = (p_1 + h, p_2 - h, p_3, p_4, \ldots, p_N) = (\alpha p_1 + (1-\alpha)p_2, \alpha p_2 + (1-\alpha)p_1, p_3, \ldots, p_N). \quad (6.1)$$

We indeed have $p' \prec p$.

Theorem 6.3. *Let p be a non-uniform probability distribution over H. Without loss of generality, we can assume that $p_1 < p_2$. Let p' be defined by* (6.1)*. Then, for all k between 1 and $N-1$,*

$$P_p(Y_n \leq k) \geq P_{p'}(Y_n \leq k), \quad (6.2)$$

which is equivalent to

$$P_p(T_{k+1} \leq n) \leq P_{p'}(T_{k+1} \leq n). \quad (6.3)$$

Moreover, if at least $k-1$ of the values $p_3, p_4, \ldots, p_N$ are non-zero, then

$$P_p(Y_n \leq k) > P_{p'}(Y_n \leq k), \quad n = k+1, k+2, k+3 \ldots \quad (6.4)$$

which is equivalent to

$$\left.\begin{array}{c} P_p(T_{k+1} \leq n) < P_{p'}(T_{k+1} \leq n), \\ n = k+1, k+2, k+3 \ldots \end{array}\right\} \quad (6.5)$$

where p' is defined by (6.1).

Proof. Can be found in annex C in supplementary electronic material. ■

We can state the following corollaries.

Corollary 6.4. *Let $p = (p_1, p_2, \ldots, p_N)$ and $q = (q_1, q_2, \ldots, q_N)$ be risk distributions on $H = \{1, 2, \ldots, N\}$. If $p \prec q$, then, for every $n \geq 1$ and every $k \geq 1$*

$$P_p(Y_n \leq k) \leq P_q(Y_n \leq k) \quad (6.6)$$

is satisfied and

$$P_p(T_{k+1} \leq n) \geq P_q(T_{k+1} \leq n).$$

Furthermore, if the distributions p and q are actually different, meaning that they do not differ only by a permutation, then the preceding inequalities are strict, except in trivial cases. More precisely, denoting by j the number of non zero p_i values (and remarking that the number of non-zero q_i values is at most j), we have

— *if $k \geq n$ or $k \geq j$, then*

$$P_p(Y_n \leq k) = P_q(Y_n \leq k) = 1 \quad \text{and}$$
$$P_p(T_{k+1} \leq n) = P_q(T_{k+1} \leq n) = 0;$$

— *if $n \geq 2$, $k < n$ and $k < j$, then*

$$P_p(Y_n \leq k) < P_q(Y_n \leq k) \quad \text{and}$$
$$P_p(T_{k+1} \leq n) > P_q(T_{k+1} \leq n).$$

Proof. As it is possible to go from vector q to vector p by a finite sequence of Robin Hood transfers, the corollary follows directly from theorem 6.3, which proves that each transfer decreases the quantity $P_p(Y_n \leq k)$. We just have to consider the cases in which this quantity is strictly decreased. ■

Remark 6.5. We can interpret the results obtained above in terms of the theory of Schur-convex functions. A real-valued function ϕ defined on a set $\mathcal{A} \subset \mathcal{R}^N$ is said to be Schur-convex on $\mathcal{A}$ if, for every x and y pair of elements in $\mathcal{A}$ such that $x \prec y$, the inequality $\phi(x) \leq \phi(y)$ holds. The first part of corollary 6.4 states that the map $p \to P_p(Y_n \leq k)$ is Schur-convex. This was already proved in [36], and was stated as a conjecture in [4].

Corollary 6.6. *Let $u = (1/N, 1/N, \ldots, 1/N)$ be the uniform distribution on $H = \{1, 2, \ldots, N\}$ and $p = (p_1, p_2, \ldots, p_N)$ any other risk distribution on H. Then*

$$P_u(Y_n \leq k) < P_q(Y_n \leq k), \quad k = 1, 2, \ldots, N-1,$$
$$n = k+1, k+2, \ldots P_u(T_{k+1} \leq n) > P_q(T_{k+1} \leq n),$$
$$k = 1, 2, \ldots, N-1, \quad n = k+1, k+2, \ \ldots$$

Proof. It can be clearly seen that $u = (1/N, 1/N, \ldots, 1/N)$ is majorized by any other distribution on H and the corollary follows. ■

Remark 6.7. The results obtained in corollaries 6.4 and 6.6 can be expressed in terms of a comparison of probability distributions as follows. If $p \prec q$, then relation (6.6) proves that the random variable Y_n defined on the probability space determined by p on the space of the random sets of $H = \{1, 2, \ldots, N\}$ is weakly stochastically dominated by the random variable Y_n defined on the probability space determined by q. Corollary 6.6 proves that the random variable Y_n defined on the probability space determined by the uniform distribution $u = (1/N, 1/N, \ldots, 1/N)$ is always weakly stochastically dominated by the random variable Y_n defined on the probability space determined by any other probability distribution on H.

Remark 6.8. A very recent and similar study in [39] proves inequalities (6.2) and (6.3) of theorem 6.3 through a different procedure. Our contribution offers a more elegant argument, based on use of fundamental formulae (3.2) and (3.3) in different contexts. Furthermore, the quality of the arguments allows us to obtain cases of strict inequalities.

7. A conjecture on strong dominance

In §6, we used an order relationship between random variables (or more precisely between their distributions) that can be defined formally as follows.

Definition 7.1. Let X and X' be two real random variables, defined on probability spaces (Ω, P) and (Ω', P'), respectively. We say that the random variable X weakly stochastically dominates the random variable X' if the cumulative distribution function of X' dominates the cumulative distribution function of X. That is, for any $t \in \mathbb{R}$,

$$P(X \leq t) \leq P'(X' \leq t).$$

The main result of §6 is that if $p \prec q$, then the random variable Y_n defined on the probability space (Ω, P_p) weakly stochastically dominates the random variable Y_n defined on the probability space (Ω, P_q).

A particular case of weak dominance is one in which inequalities apply not only to the cumulative distribution functions, but also to the distributions themselves. We refer to this situation as strong dominance, and we provide a formal definition of strong dominance below, for the case of discrete random variables. (A similar definition can be given for continuous random variables with densities.) In short, X strongly dominates X' if, for any small enough value d, $P(X = d) \leq P'(X' = d)$, and if for any other possible value e, $P(X = e) \geq P'(X' = e)$.

Definition 7.2. Let X and X' be two real random variables, defined on probability spaces (Ω, P) and (Ω', P'), respectively, and taking values in a denumerable set D. We say that the random variable$X$ strongly stochastically dominates the random variable X' if there is a critical value $c \in \mathbb{R}$ such that, for any $d \in D$

— if $d \leq c$, then $P(X = d) \leq P'(X' = d)$ *and*
— if $d > c$, then $P(X = d) \geq P'(X' = d)$.

It is easy to show that strong dominance implies weak dominance, but that the converse is not true. Coming back to our CCP model, we propose the following.

Conjecture. If $p \prec q$, then the random variable Y_n defined on the probability space (Ω, P_p) strongly stochastically dominates the random variable Y_n defined on the probability space (Ω, P_q).

This conjecture has been tested on various examples, but we have been able to prove it formally for only a few values of the pair (n, N), namely for $n = 2$ or 3 and any N, and for $n = 4$ and $N \leq 5$.

In applications, strong dominance reinforces weak dominance. It gives more precise statements concerning the relative probabilities that a given number of hosts are parasitized after a given number of eggs laid, for different risk distributions.

8. Illustrative examples

In this section, we show graphically the relationships satisfied among the distribution functions of random variables Y_n as well as the distribution functions of random variables T_k, when their corresponding risk distributions are able to be compared by majorization.

Figures 2 and 3 represent very different situations in terms of risk distribution over hosts. The risk distributions are majoring each other in figure 2, whereas this is not the case for figure 3.

The distribution functions of five variables Y_n are represented in figure 2*a*. They correspond to five different risk distributions, p_1, p_2, p_3, p_4 and p_5, satisfying $p_1 \prec p_2 \prec p_3 \prec p_4 \prec p_5$. These are distributions on the set $\{1, 2, \ldots, 12\}$ (so $N = 12$), p_1 is the uniform distribution, $p_i = (1/(10i+2), \ldots, 1/(10i+2), (10(i-1)+1)/(10i+2))$ for $i = 2$ and 3, and $p_i = (1/(45(i-3)+12), \ldots, 1/(45(i-3)+12), 1/(45(i-3)+12), (45(i-3)+1)/(45(i-3)+12))$ for $i = 4$ and 5. We have also used $n = 12$, and it can be observed that $P_{p_i}(Y_{12} \leq k) < P_{p_{i+1}}(Y_{12} \leq k)$, with $k = 1, 2, \ldots, 11$, $i = 1, 2, 3, 4$. This panel shows that, for example, the probability to observe no more than five parasitized hosts with 12 draws is nearly one for the distribution drawn in red, and nearly zero for the distribution in black.

The distribution functions of 10 variables T_k are represented in figure 2*b*,*c*. The number of hosts, N, and the risk distributions are the same in both cases; $N = 10$, p_1 is the uniform distribution and $p_i = (1/5(i+1), \ldots, 1/5(i+1), i/5(i+1), (4(i-1)+1)/(5(i+1)))$, for $i = 2, 3, \ldots, 10$. For these risk distributions, $p_1 \prec p_2 \prec \ldots p_9 \prec p_{10}$ is satisfied. In figure 2*b*, $k = 6$ and the values of n lie between 6 to 50. In figure 2*c*, $k = 9$ and the values of n lie from 9 to 100. It can be seen that $P_{p_i}(T_k \leq n) > P_{p_{i+1}}(T_k \leq n)$, for $n = k, k+1, \ldots$ $i = 1, 2, \ldots 9$, in both graphics. For example, figure 2*b* shows that the probability to observe at least six parasitized hosts when sampling 10 hosts is nearly 0.8 for the distribution in black, whereas at least 50 hosts have to be sampled in order to obtain the same probability in the situation described by the green curve in the bottom of the panel. Figure 2*c* shows a similar case for nine parasitized hosts. One clearly observes that the ordering of the distributions of random variables Y_n and T_k follows the ordering of the risk distributions, described above. This is very different if the risk distributions do not have any simple ordering among themselves, as

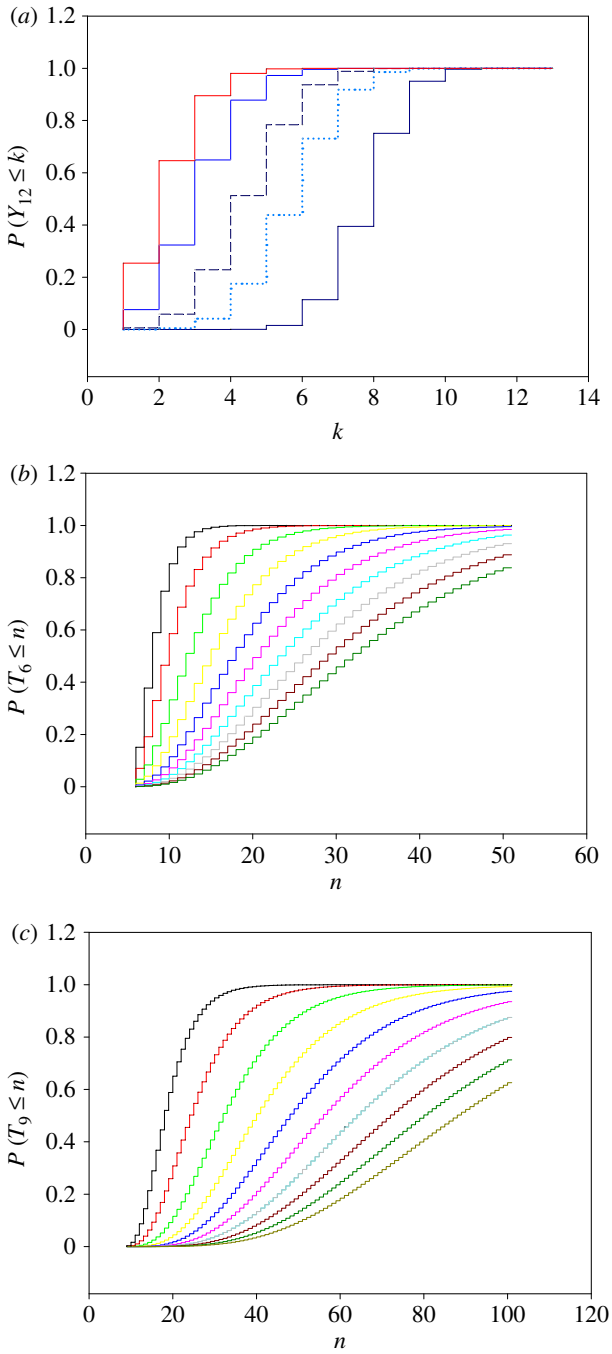

Figure 2. Numerical examples with ordered risk distributions. (*a*) Distribution functions of five variables Y_{12} corresponding to five different risk distributions, $p_1, p_2, \ldots, p_5$, satisfying $p_1 \prec p_2 \prec p_3 \prec p_4 \prec p_5$. It can be observed that $P_{p_i}(Y_{12} \leq k) < P_{p_{i+1}}(Y_{12} \leq k)$, for $k = 1, 2, \ldots, 11$, $i = 1, 2, 3, 4$. (*b*) Distribution functions of ten variables T_6 corresponding to ten risk distributions, $p_1, p_2, \ldots, p_{10}$ satisfying $p_1 \prec p_2 \prec \ldots p_9 \prec p_{10}$. (*c*) Distribution functions of ten variables T_9 corresponding to the same previous risk distributions. In panels (*b*) and (*c*), it can be observed that $P_{p_i}(T_k \leq n) > P_{p_{i+1}}(T_k \leq n)$, for $n = k, k+1, \ldots$ $i = 1, 2, \ldots 9$.

shown in figure 3. Figure 3 compares distribution functions of random variables T_k corresponding to two unrelated risk distributions p and q, i.e. neither $p \prec q$ nor $q \prec p$. Thus, these distribution functions act in different ways depending on the value of k. We include three different graphics, each bearing two curves. These curves are the distribution functions of two random variables T_k. The risk distributions associated with these random variables are, in the three graphics, $p = \left(\frac{3}{85}, \frac{3}{85}, \frac{3}{85}, \frac{3}{85}, \frac{3}{85}, \frac{12}{85}, \frac{13}{85}, \frac{13}{85}, \frac{13}{85}, \frac{19}{85}\right)$ and

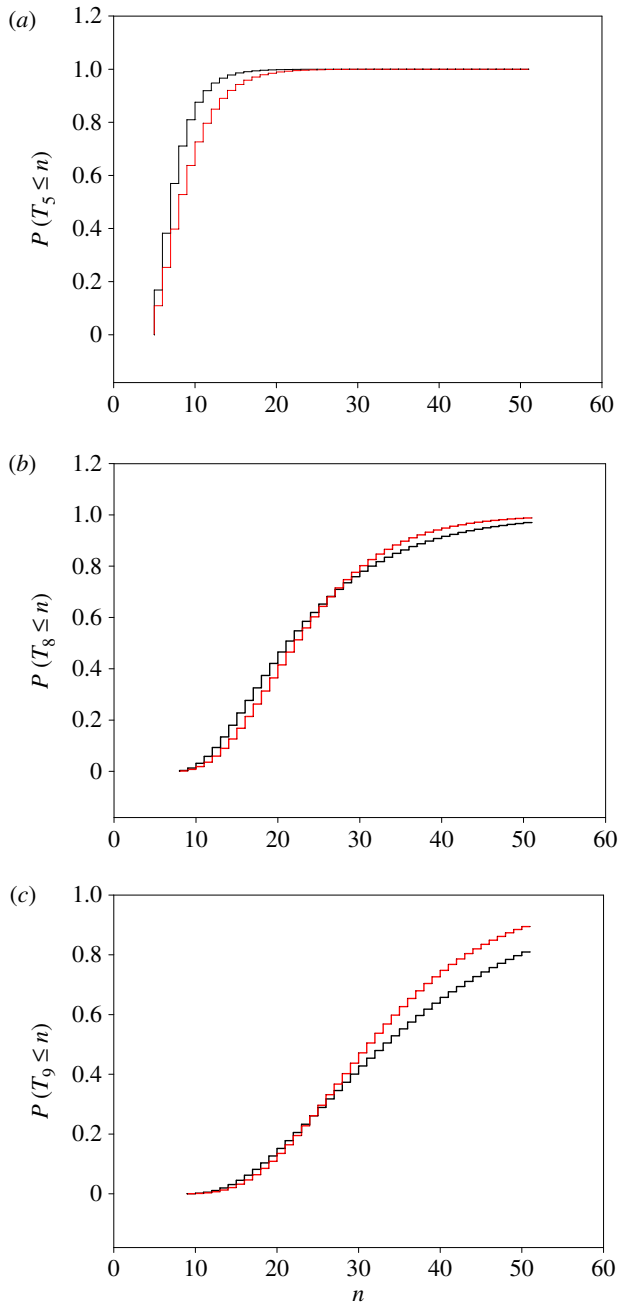

Figure 3. Numerical examples with risk distributions without first-order stochastic dominance. Comparison of distribution functions of random variables T_k corresponding to two unrelated risk distributions, i.e. neither $p \prec q$ nor $q \prec p$, to show how these distribution functions act in different ways depending on the value of k.

$q = \left(\frac{3}{81}, \frac{4}{81}, \frac{4}{81}, \frac{4}{81}, \frac{4}{81}, \frac{5}{81}, \frac{5}{81}, \frac{5}{81}, \frac{15}{81}, \frac{32}{81}\right)$. In figure 3*a*, $k = 5$, and the probability that at least five hosts have been parasitized after n draws is always less for the distribution in red than for the one in black. It does not occur in figure 3*b*,*c*, where $k = 8$ and 9, respectively. In these cases, the probabilities that at least k hosts have been parasitized after n draws are lower for the red distribution than for the black one when the number of draws, n, is small, but are reversed when n is larger.

9. Applying our approach to a real host–parasitoid system

The population dynamics of host–parasitoid systems have often been theoretically studied in terms of urn processes,

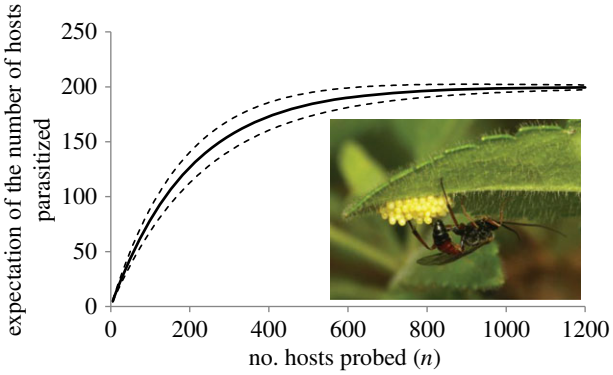

Figure 4. Theory applied to the specific biological interaction between an egg parasitoid, *Hyposoter horticola* (Hymenoptera: Ichneumonidae), and its butterfly host, *Melitaea cinxia* (Lepidoptera: Nymphalidae). The number of hosts probed to attain a given level of parasitism is increasing with increasing percentage of the parasitism rate. The 90% variability bands were computed using the Tchebycheff inequality.

as explained in Introduction. The otherwise quite vast literature on parasitoid behaviour seems however nearly void of real datasets available for applying urn processes. The single reference we identified containing all needed information, namely both the host number identity and the handling behaviour of the wasp over time, is by Simmonds [40], but the datasets presented there are too sparse to enable modelling. Van Lenteren presents in [41] figures with nearly complete information, save for host identity. We therefore use the biological interaction between the butterfly *Melitaea cinxia* (Lepidoptera: Nymphalidae) and its ichneumonid parasitoid *Hyposoter horticola* (Hymenoptera: Ichneumonidae) as an example (see inset in figure 4). The wasp females search for and find the butterfly egg clusters and attack them. We chose this system for several reasons. First, we know more on this system in the wild than on many others [3,42–44]. Second, butterflies lay their eggs in clusters, enabling a clear identification of the number of hosts available. Third, a female leaves a clutch after having parasitized only a third of the eggs. The puzzle of leaving most of a resource unexploited is explained in this case by the avoidance of superparasitism [26]. Indeed, as only one larva survives in a host, superparasitism should be avoided when encountering previously parasitized hosts. In reality, it was observed to occur occasionally, with a probability of some 20%. Our model is silent about the fate of superparasitized hosts. Fourth, the wasp is 'making haphazard passes across the cluster' [26, p. 543] and does not show any systematic way of searching, enabling us to assume that the hypothesis of uniform distribution of risk at the cluster level is valid. The authors state that 'because about 30% of each roughly 200-egg host cluster is parasitized, the wasp must probe, on average, more than 60 eggs per cluster'. In the following, we use the theory developed in this paper to make the above statement more precise and more generic.

Our approach defines the time axis by the number of successive draws. Handling time, a key parameter in most models of host–parasitoid systems, is therefore included in the concept of a draw, i.e. time is not measured in seconds or minutes. Assuming no parasitism at the arrival of the wasp on the 200-eggs cluster, the expected number of probes necessary to have 60 hosts parasitized, $E(T_{60})$, is 72, according to equation (4.5). The number of necessary draws to attain a high percentage of parasitism is increasing with

that level of parasitism and explodes when approaching 80% parasitism and more: it needs nearly 1200 draws to parasitize 199 hosts. This is the 'diminishing return' explanation given by the authors for *Hyposoter*: a foraging female might, due to incomplete knowledge of parasitism status of the hosts, superparasitize some. The number of expected parasitized hosts, $E(Y_n)$, as a function of the number of hosts probed is given in figure 4. Making inverse inference from Y_n to T_k is not immediate, because we deal with stochasticity: for a single trajectory of the whole dynamics, the map $k \to T_k$ is a well-defined generalized inverse of the map $n \to Y_n$, but this fact has no direct translation on the link between the maps $k \to E(T_k)$ and $n \to E(Y_n)$. In other words, the strong correspondence between the distribution functions of random variables Y_n and T_k has no direct translation on their expectations. A similar problem often encountered by biologists is inverse regression. The behaviour of the variance is interesting: it is very small early on, as each sampled host will most likely be unparasitized (the variance is actually null for first draw), maximal at mid-range and small again at very large numbers of parasitized hosts.

The eggs of a single cluster are building 'mounts', and some hosts might be more accessible than others. While Montovan *et al*. [26] have checked the uniformity of parasitism according to depth in the mount and found no gradient, risk of parasitism is known to be varying with accessibility in other host–parasitoid systems (see references in [26]). Let us assume that a single host egg has a probability α of being parasitized and the others have probability β of being parasitized. This is the extreme case of the situation studied in §5.2. From (5.4), we obtain the following expressions for $E(T_{k,N_1})$:

(1) If the host with probability α of being parasitized does not belong to the set H_1, then

$$E(T_{k,N_1}) = \frac{1}{\beta}\sum_{j=0}^{k-1}\frac{1}{N_1 - j}.$$

(2) If the host with probability α of being parasitized belongs to the set H_1, then

$$\begin{aligned} E(T_{k,N_1}) = {} & (N_1-1)^{(k)}\frac{\beta^k}{\prod_{i=1}^{k}(\alpha+(N_1-i)\beta)}\sum_{i=1}^{k}\frac{1}{\alpha+(N_1-i)\beta} \\ & + (N_1-1)^{(k-1)}\sum_{j=1}^{k-1}\left(\frac{\alpha\beta^{k-1}}{\prod_{i=1}^{j}(\alpha+(N_1-i)\beta)\prod_{i=j}^{k-1}(N_1-i)\beta}\sum_{i=1}^{j}\frac{1}{\alpha+(N_1-i)\beta}+\sum_{i=j}^{k-1}\frac{1}{(N_1-i)\beta}\right) \\ & + (N_1-1)^{(k-1)}\frac{\alpha\beta^{k-1}}{\prod_{i=1}^{k}(\alpha+(N_1-i)\beta)}\sum_{i=1}^{k}\frac{1}{\alpha+(N_1-i)\beta}. \end{aligned} \tag{9.1}$$

Now, let us assume, for the purpose of illustration, that a single host egg, out of the 200, has a risk of being parasitized 10 times higher than all the others. For this very specific case, we obtain by applying (9.1) the value of $E(T_{60})$ of 73 to have 60 hosts parasitized, an insignificant increase over the situation of uniform risk distribution of all hosts. If this host has now a risk of parasitism 100 times higher than all the others, $E(T_{60})$ increases markedly, to 105. Two observations can be made on the basis of these computations. First, it seems difficult to inversely infer the risk distribution among hosts on the basis of some characteristics of the random variables T_k or Y_n. It is in fact nearly impossible to do so if we do not know N, k, and the fact that there is only one host different from the others. Theoretical developments would be valuable to enable such inverse inference, of high importance in practical terms. Second, the increased number of draws with increased risk heterogeneity reflects the shadowing effect mentioned earlier, in which this single host with higher risk of parasitism 'protects' the others, by being sampled more often.

10. Conclusion

Urn sampling processes are well understood when items have a uniform probability to be picked up [45]. This simplest model and its implications, such as the Poisson distribution of draws per sampled unit for large populations, have been frequently used in ecology and evolution. The theoretical developments in this area are however not complete, as some fundamental statistics of practical importance, such as the variance of T_k, have not been worked out in the probability theory literature yet. This area thus represents a worthwhile probabilistic development. Dealing with non-uniform distributions of risk is more difficult, both because the mathematics are indeed more elaborate and because the outcome is highly sensitive to variations of the risk distributions. Distributions of risk observed in real, natural systems are however highly non-uniformly distributed and often markedly skewed [28,29], hence the need to expand the research field into this direction. The relationship between risk distribution and population dynamics remains complex to understand, constrained by strong assumptions about timing of events and depending on the entire distribution of risk rather than the variance only (see [30] and the following flurry of publications). As a consequence, the assessment of the influence of different distributions of risk of parasitism on the population dynamics of host–parasitoid systems relied so far only on extensive simulations [32]. The next challenge is thus to integrate biased urn sampling theory with their population dynamics models. How the stability properties of the interaction are influenced by the dominance

properties of various distributions of risk of parasitism can then be assessed in generic terms.

Authors' contributions. J.C. framed the question, N.Z., E.L. and M.J.F.S. carried out the bulk of the mathematical reasoning, E.L. proposed the new conjecture, P.Z. computed the illustrative examples and J.C. and N.Z. worked out the biological example. The first draft was written by N.Z., with amendments by J.C. and E.L.

Funding. N.Z. and M.J.F.S. acknowledge the financial support of the Fundación Seneca of the Comunidad Autónoma de la Región de Murcia, project no.19320/IP/14.

Acknowledgments. We thank Dr van Nouhuys and Chicago University Press for the use of the photograph of parasitoid from [26]. N.Z. is also grateful to the University François-Rabelais of Tours, for its support and hospitality.

References

1. Feller V. 1968 *An introduction to probability theory and its applications*, vol. 1. New York, NY: John Wiley & Sons.
2. Boneh A, Hofri M. 1989 The coupon-colector problem revisited. Computer Science Technical Report. Paper 807. See http://docs-lib.purdue.edu/cstech/807.
3. Flajolet P, Gardy D, Thimonier L. 1992 Birthday paradox, coupon collectors, caching algorithms and self-organizing search. *Discrete Appl. Math.* **39**, 207–229. (doi:10.1016/0166-218X(92)90177-C)
4. Anceaume E, Busnel Y, Sericola B. 2015 New results on a generalized coupon collector problem using Markov chains. *J. Appl. Probab.* **52**, 405–418. (doi:10.1017/S0021900200012547)
5. MacArthur RH. 1957 On the relative abundance of bird species. *Proc. Natl Acad. Sci. USA* **43**, 293–295. (doi:10.1073/pnas.43.3.293)
6. Simpson EH. 1949 Measurement of diversity. *Nature* **163**, 688. (doi:10.1038/163688a0)
7. Bunge J, Fitzpatrick M. 1993 Estimating the number of species: a review. *J. Am. Stat. Assoc.* **88**, 364–373. (doi:10.2307/2290733)
8. Huillet T, Paroissin C. 2009 Sampling from Dirichlet partitions: estimating the number of species. *Environmetrics* **20**, 853–876. (doi:10.1002/env.977)
9. Dennehy JJ. 2009 Bacteriophages as model organisms for virus emergence research. *Trends Microbiol.* **17**, 450–457. (doi:10.1016/j.tim.2009.07.006)
10. Neal P, Moriary J. 2009 Sampling efficiency and biodiversity. Research Report 9, Probability and Statistics Group School of Mathematics. University of Manchester.
11. Daley DJ, Gani J, Gani JM. 2001 *Epidemic modelling: an introduction*. Cambridge, UK: Cambridge University Press.
12. Lloyd-Smith JO, Schreiber SJ, Kopp PE, Getz WM. 2005 Superspreading and the effect of individual variation on disease emergence. *Nature,* **438**, 355–359. (doi:10.1038/nature04153)
13. Keeling MJ, Rohani P. 2008 *Modelling infectious diseases in humans and animals*. Princeton, NJ: Princeton University Press.
14. Hernández-Suárez CM, Mendoza-Cano O. 2009 Applications of occupancy urn models to epidemiology. *Math. Biosci. Eng.* **6**, 509–520. (doi:10.3934/mbe.2009.6.509)
15. Bailey LL, MacKenzie DI, Nichols JD. 2014 Advances and applications of occupancy models. *Methods Ecol. Evol.* **5**, 1269–1279. (doi:10.1111/2041-210X.12100)
16. Hernández-Suárez CM, Hiebeler D. 2012 Modeling species dispersal with occupancy urn models. *Theor. Ecol.* **5**, 555–565. (doi:10.1007/s12080-011-0147-8)
17. Kershenbaum A, Freeberg TM, Gammon DE. 2015 Estimating vocal repertoire size is like collecting coupons: a theoretical framework with heterogeneity in signal abundance. *J. Theor. Biol.* **373**, 1–11. (doi:10.1016/j.jtbi.2015.03.009)
18. Ewens WJ. 1972 The sampling theory of selectively neutral alleles. *Theor. Popul. Biol.* **3**, 87–112. (doi:10.1016/0040-5809(72)90035-4)
19. Donnelly P. 1986 Partition structures, Polya urns, the Ewens sampling formula, and the ages of alleles. *Theor. Popul. Biol.* **30**, 271–288. (doi:10.1016/0040-5809(86)90037-7)
20. Doumas AV. 2015 How many trials does it take to collect all different types of a population with probability *p*? *J. Appl. Math. Bioinf.* **5**, 1–14.
21. Dixon CJ. 2006 A means of estimating the completeness of haplotype sampling using the Stirling probability distribution. *Mol. Ecol. Notes* **6**, 650–652. (doi:10.1111/j.1471-8286.2006.01411.x)
22. Tenaillon O, Rodríguez-Verdugo A, Gaut RL, McDonald P, Bennett AF, Long AD, Gaut BS. 2012 The molecular diversity of adaptive convergence. *Science* **335**, 457–461. (doi:10.1126/science.1212986)
23. Vandewalle K, Festjens N, Plets E, Vuylsteke M, Saeys Y, Callewaert N. 2015 Characterization of genome-wide ordered sequence-tagged *Mycobacterium* mutant libraries by Cartesian pooling-coordinate sequencing. *Nat. Commun.* **6**, 7106. (doi:10.1038/ncomms8106)
24. Fiske WF. 1910 Superparasitism: an important factor in the natural control of insects. *J. Econ. Entomol.* **3**, 88–97. (doi:10.1093/jee/3.1.88)
25. Thompson WR. 1924 La théorie mathématique de l'action des parasites entomophages et le facteur du hasard. *Ann. Fac. Sci. Marseille* **2**, 69–89.
26. Montovan KJ, Couchoux C, Jones LE, Reeve HK, van Nouhuys S. 2015 The puzzle of partial resource use by a parasitoid wasp. *Am. Nat.* **185**, 538–550. (doi:10.1086/680036)
27. Hemerik L, Van der Hoeven N, van Alphen JJ. 2002 Egg distributions and the information a solitary parasitoid has and uses for its oviposition decisions. *Acta Biotheor.* **50**, 167–188. (doi:10.1023/A:1016543310896)
28. Hassell M. 2000 *The spatial and temporal dynamics of host–parasitoid interactions*. Oxford, UK: Oxford University Press.
29. Murdoch WW, Briggs CJ, Nisbet RM. 2013 *Consumer–resource dynamics* (*MPB-36*). Princeton, NJ: Princeton University Press.
30. May RM. 1978 Host–parasitoid systems in patchy environments: a phenomenological model. *J. Anim. Ecol.* **47**, 833–844. (doi:10.2307/3674)
31. Ives AR, Schooler SS, Jagar VJ, Grbic M, Settle WH. 1999 Variability and parasitoid foraging efficiency: a case study of pea aphids and *Aphidius ervi*. *Am. Nat.* **154**, 652–673. (doi:10.1086/303269)
32. Singh A, Murdoch WW, Nisbet RM. 2009 Skewed attacks, stability, and host suppression. *Ecology* **90**, 1679–1686. (doi:10.1890/07-2072.1)
33. Casas J. 1989 Foraging behaviour of a leafminer parasitoid in the field. *Ecol. Entomol.* **14**, 257–265. (doi:10.1111/j.1365-2311.1989.tb00954.x)
34. Casas J, Swarbrick S, Murdoch WW. 2004 Parasitoid behaviour: predicting field from laboratory. *Ecol. Entomol.* **29**, 657–665. (doi:10.1111/j.0307-6946.2004.00647.x)
35. Marshall AW, Olkin I, Arnold BC. 2011 *Inequalities: theory of majorization and its applications*, 2nd edn. Springer Series in Statistics. New York, NY: Springer.
36. Wong CK, Yue PC. 1973 A majorization theorem for the number of distinct outcomes in N independent trials. *Discrete Math.* **6**, 391–398. (doi:10.1016/0012-365X(73)90070-8)
37. Arnold B. 1987 *Majorization and the Lorenz order: a brief introduction*. Berlin, Germany: Springer.
38. Muirhead RF. 1902 Some methods applicable to identities and inequalities of symmetric algebraic functions of n letters. *Proc. Edinb. Math. Soc.* **21**, 144–162. (doi:10.1017/S001309 150003460X)
39. Anceaume E, Busnel Y, Shulte-Geers E, Sericola B. 2016 Optimization results for a generalized coupon collector problem. *J. Appl. Probab.* **53**, 622–629. (doi:10.1017/jpr.2016.27)
40. Simmonds FJ. 1943 The occurrence of superparasitism in *Nemeritis canescens*. *Grav. Rev. Can. Biol.* **2**, 15–58.
41. van Lenteren JC. 1975 The development of host discrimination and the prevention of superparasitism in the parasite *Pseudeucoila bochei* Weld (Hym.: Cynipidae). *Neth. J. Zool.* **26**, 1–83. (doi:10.1163/002829676X00055)

42. van Nouhuys S. 2016 Diversity, population structure, and individual behaviour of parasitoids as seen using molecular markers. *Curr. Opin. Insect Sci.* **14**, 94–99. (doi:10.1016/j.cois.2016.02.006)
43. van Nouhuys S, Ehrnsten J. 2004 Wasp behavior leads to uniform parasitism of a host available only a few hours per year. *Behav. Ecol.* **15**, 661–665. (doi:10.1093/beheco/arh059)
44. van Nouhuys S, Kaartinen R. 2008 A parasitoid wasp uses landmarks while monitoring potential resources. *Proc. R. Soc. B* **275**, 377–385. (doi:10.1098/rspb.2007.1446)
45. Kotz S, Balakrishnan N. 1997 Advances in urn models during the past two decades. In *Advances in combinatorial methods and applications to probability and statistics* (ed. N Balakrishnan), pp. 203–257. Boston, MA: Birkhäuser. (doi:10.1007/978-1-4612-4140-9_14)