

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

Noemi Zoroa, Emmanuel Lesigne, José Fernandez-Saez, P Zoroa, Jérome Casas

▶ **To cite this version:**

Noemi Zoroa, Emmanuel Lesigne, José Fernandez-Saez, P Zoroa, Jérome Casas. The coupon collector urn model with unequal probabilities in ecology and evolution. Journal of the Royal Society Interface, 2017, 14, 10.1098/rsif.2016.0643 . hal-01354100

HAL Id: hal-01354100
https://univ-tours.hal.science/hal-01354100

Submitted on 18 Aug 2016

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

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

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

August 15, 2016

[1]Departamento de Estadística e Investigación Operativa, Facultad de Matematicas, Universidad de Murcia, 30071, Murcia, SPAIN. [2]Université François-Rabelais, CNRS, LMPT UMR7350, Tours, FRANCE. [3]Université de Tours & Institut Universitaire de France Institut de Recherche en Biologie de l'Insecte, IRBI UMR CNRS 7261 37200, Tours, FRANCE

Abstract

The sequential sampling of populations with unequal probabilities and with replacement in a closed population is a recurrent problem in ecology and evolution. 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 color 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 speed 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 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 to two different distributions, as generally assumed in ecological studies.

Keywords: Coupon collector's problem; parasitoid; stochastic dominance; strong dominance; ecology.

2010 Mathematics Subjet Classification: 60C05, 60E15, 92B05

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 (Feller, 1968). Rapid advances have been made in this field (Boneh and Hofri, 1989; Flajolet et al., 1992; Anceaume et al., 2015), but they have gone largely unnoticed 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 (McArthur, 1957; Simpson, 1949). This field could greatly benefit from the latest advances in the CCP (Fitzpatrick, 1993; Huillet and Paroissin, 2009). Related problems deal with the relative abundance of species from a community containing many species (Dennehy, 2009), or the sampling effort required to achieve a particular level of coverage (Neal and Moriary, 2009). Increases in the number of new hosts being infected or superinfected are a topic of great importance in population dynamics and epidemiology (Daley et al., 2001; Lloyd-Smith et al., 2005; Keeling and Rohani, 2008). 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 Bailey et al., 2013 and the paper Hernandez-Suarez and Hiebeler, 2011). 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 behavioral indicator of the complexity of the vocal communication system in birds and mammals (Kershenbaum et al., 2015). In genetics and evolution, the coupon collector problem has been recognized as such only occasionally, despite these fields having generated some of the most elegant theorems and uses of other urn processes (Ewens, 1972; Donelly, 1986; Doumas, 2015). Indeed, the coupon collector problem has been used in the context of exhaustive haplotype sampling in phylogeography, (Dixon, 2006), determining the number of beneficial mutations as a function of sequence lines (Tenaillon et al., 2012), and estimations of the size of the library required to target a particular percentage of the non-essential genome displaying a given property (Vandewalle et al., 2015), 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 'color' by parasitizing them. In work beginning more than a century ago (Fiske, 1910; Thompson, 1924), 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 (Montovan et al., 2015). 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 (Hemerick et al., 2002). Aggregation is interpreted as the result of heterogeneity in the risk of being found, due to differences in location, accessibility, appearance, color, developmental stage or any other trait (Hassell, 2000; Murdoch et al., 2013). The risk distribution greatly influences the stability of the host-parasitoid system and has been widely studied (May, 1978; Ives et al., 1999; Singh et al., 2009).

All these models focus on the distribution of eggs over the entire population of hosts, after a given time or a given number of draws (Figure1). However, the use of this distribution greatly decreases the amount of information available, as it collapses individual host histories. Parasitism is a multinomial process (Figure1), 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 the present paper, we aimed to model parasitism as a multinomial urn process over time and we study the speed of parasitization (Figure1). We consider host encounters followed by oviposition without discrimination. The parasitism process described above can be considered as a coupon collector problem. 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 coupon collector problem, 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

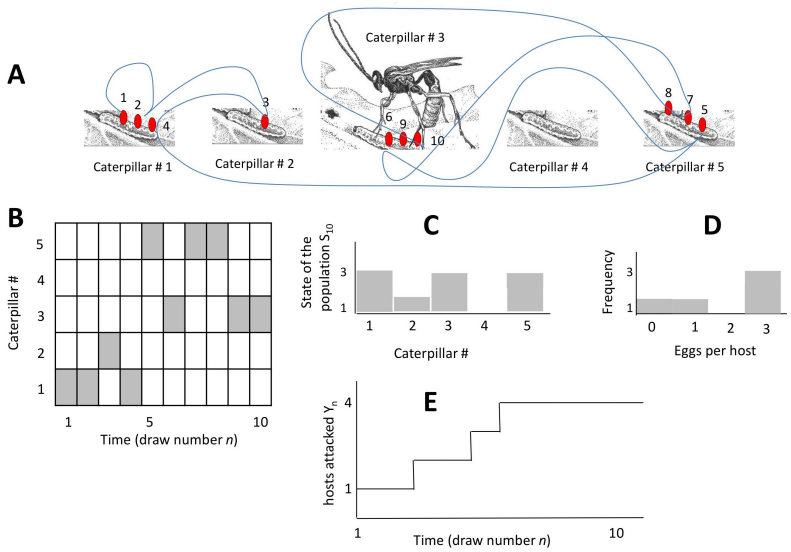


Figure 1. Attacks of caterpillar larvae hosts by parasitic wasps as an urn process in discrete time n. A wasp oviposit 10 eggs among five hosts ($n = 10$) (A). The outcome of the fundamental multinomial process (B) is summarized in the marginal distribution of the number of eggs per individual host at 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)

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 Section 2, we define a succession S_n, $n = 1, 2, ...$, 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 h-th component representing the number of times that host h has been attacked by a parasitoid between times 1 and n. In Section 3 we define the random variables Y_n, $n = 1, 2, ...$, 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 Section 4. In Section 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 Section 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. We highlight this finding with numerical examples in Section 7 and propose a conjecture on strong stochastic dominance in Section 8.

2 Modelling parasitism as an urn process

We assume a finite population of hosts, constant for the entire duration of the experiment. 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 due 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,...,p_N or $(p_1, p_2, ..., p_N)$ is the risk distribution for the set of hosts $H = \{1, 2, ..., N\}$.

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 $\{\omega = i_1 i_2 i_3 \ldots i_n j_{n+1} j_{n+2} \cdots : \text{ for some } j_k \text{in } H, k > n\}$ is $p_{i_1} p_{i_2} \ldots p_{i_n}$. When necessary, the vector $(p_1, p_2, ..., p_N)$ 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,

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

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 $i = 1, 2, ..., N$, then S_n takes the value $(r_1, r_2, ..., r_N)$. This variable has a multinomial distribution with parameters n, p_1,p_2,...,p_N, that is, for every integers r_1, r_2,..., r_N, $0 \leq r_i \leq n$, $r_1 + r_2 + ... + r_N = n$,

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

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

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

with $\quad 0 \leq r_{i_j} \leq n, \ j = 1, 2, ..., h, \ r_{i_1} + r_{i_2} + ... + r_{i_h} \leq n, \ q_{i_1 i_2 ... i_h} = 1 - \sum_{j=1}^{h} p_{i_j}$.

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,

$$P(S_{nh} = r) = \frac{n!}{r!(n-r)!} p_h^r q_h^{n-r}, r = 0, 1, 2, ..., n \tag{2.4}$$

$$\text{where } q_h = 1 - p_h = \sum_{i \neq h} p_i.$$

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

The expected value and variance of this random variable are

$$E(S_{nh}) = np_h, \quad Var(S_{nh}) = np_h(1 - p_h).$$

Let $(e_1, e_2, ..., e_N)$ denote the canonical base of the space $\mathbb{R}^N$. 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 behavior 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 study this random variable obtaining expressions for: its probability mass function (3.3), distribution function (3.5), expectation (3.6) and variance (3.7).

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

$$x^{(h)} = x(x-1)(x-2)...(x-h+1), \quad x^{(0)} = 1,$$

$$\binom{x}{h} = \frac{x^{(h)}}{h!} = \frac{x(x-1)...(x-h+1)}{h!}, \text{ and } \binom{x}{0} = 1. \tag{3.1}$$

The distribution and the distribution function of Y_n have been obtained in previous studies, see Anceaume et al. (2015) .

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

$$P(Y_n = k) =$$

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

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

$$+ (-1)^{k-1} \begin{pmatrix} N-1 \\ N-k \end{pmatrix} \sum_{\{j\} \subset \{1,2,...,N\}} p^n_j. \tag{3.2}$$

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

$$P(Y_n = k) = \sum_{A \subset H, |A| \leq k} (-1)^{k-|A|} \binom{N-|A|}{k-|A|} p_A^n, \quad \text{for } 0 \leq k \leq N, \text{ and } k \leq n. \tag{3.3}$$

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

Let us now consider the distribution function of Y_n,

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

$$\sum_{A \subset H, |A| \leq k} \left(\sum_{i=0}^{k-|A|} (-1)^i \binom{N-|A|}{i} \right) p_A^n. \tag{3.4}$$

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

$$P(Y_n \leq k) = \sum_{A \subset H, |A| \leq k} (-1)^{k-|A|} \binom{N-|A|-1}{k-|A|} p_A^n, \quad k = 1, 2, ..., N. \tag{3.5}$$

A similar expression can be seen in Anceaume et al. (2015) . From (3.5) 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:

$$E(Y_n) = N - m_{N-1}^{[n]} = N - \sum_{i=1}^{N} (1 - p_i)^n. \tag{3.6}$$

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

$$Var(Y_n) = 2 \sum_{1 \leq i < j \leq N} (1 - p_i - p_j)^n + \sum_{i=1}^{N} (1 - p_i)^n \left(1 - \sum_{i=1}^{N} (1 - p_i)^n \right). \tag{3.7}$$

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., are 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 Section 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 $H_1 = \{1, 2, ...N_1\}$ without lost 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. The case $H_1 = H$ 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 Boneh and Hofri (1989), an expression is proposed for the particular case in which $k = N_1 = N$. The expression obtained here is more general.

Let $i_1, i_2, ..., i_k$ be distinct elements of H_1. Let $D_{i_1 i_2 ... i_k}$ 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 $i_1, i_2, ..., i_k$ and are parasitized in the precise order $i_1, i_2, ..., i_k$. 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 $p = \sum_{i \in H_1} p_i$, $q = 1 - p = \sum_{i \in H_2} p_i$. Then

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

$$P(\text{second host of } H_1 \text{ parasitized is } i_2 \mid \text{first host of } H_1 \text{ parasitized was } i_1) \ ...$$

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

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,

$$P(D_{i_1 i_2 ... i_k}) = \frac{\prod_{j=1}^{k} p_{i_j}}{p(p - p_{i_1})(p - p_{i_1} - p_{i_2})...(p - \sum_{j=1}^{k-1} p_{i_j})}. \tag{4.2}$$

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 $i_1 i_2 ... i_{N_1} \neq j_1 j_2 ... j_{N_1}$ 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,

$$E(T_{k,N_1}) = \sum_{(i_1 i_2 ... i_k) \in \Pi_k} E(T_{k,N_1} | D_{i_1 i_2 ... i_k}) P(D_{i_1 i_2 ... i_k}). \tag{4.3}$$

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

$$T_{k,N_1} = X_1 + 1 + X_2 + 1 + ... + X_k + 1 = X_1 + X_2 + ... + X_k + k. \tag{4.4}$$

and therefore

$$E(T_{k,N_1} | D_{i_1 i_2 ... i_k}) = \sum_{h=1}^{k} E(X_h | D_{i_1 i_2 ... i_k}) + k \tag{4.5}$$

but

$$E(X_h|D_{i_1 i_2 ... i_k}) = E\left(X_h|\text{already parasitized hosts are those of } H_2 \text{ and } i_1 i_2 ... i_{h-1}\right) \tag{4.6}$$

One direct application of (4.1) would then be:

$$E(X_h|D_{i_1 i_2 ... i_k}) = \frac{q + \sum_{j=1}^{h-1} p_{i_j}}{p - \sum_{j=1}^{h-1} p_{i_j}} \quad \text{for } h = 1, 2, ..., k. \tag{4.7}$$

From (4.5) and (4.7)

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

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

$$\frac{1}{1-q} + \frac{1}{1 - q - p_{i_1}} + ... + \frac{1}{1 - q - \sum_{j=1}^{k-1} p_{i_j}}, \tag{4.8}$$

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 (4.3), (4.2) and (4.8) we can state the following:

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

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

$$\frac{\prod_{j=1}^{k} p_{i_j}}{p(p - p_{i_1})(p - \sum_{j=1}^{2} p_{i_j})...(p - \sum_{j=1}^{k-1} p_{i_j})}, \tag{4.9}$$

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

Thus, $E(T_{k,N_1})$ given by (4.9) 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, ..., N_1, that is $N_1^{(k)} = N_1(N_1 - 1)...(N_1 - k + 1)$. 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, ...,p_{N_1} be real numbers satisfying $p_1 \geq p_2 \geq ... \geq p_{N_1}$. Then, the maximum of $E(T_{k,N_1}|D_{i_1 i_2 ... i_k})$ defined by (4.8) over all possible choices of the k-subsets $\{i_1, i_2, ..., i_k\}$ of H_1 is

$$E(T_{k,N_1}|D_{1,2,...,k}) = \frac{1}{\sum_{i=1}^{N_1} p_i} + \frac{1}{\sum_{i=2}^{N_1} p_i} + ... + \frac{1}{\sum_{i=k}^{N_1} p_i}, \tag{4.10}$$

and the minimum is

$$E(T_{k,N_1}|D_{N_1,N_1-1,...,N_1-k+1}) = \frac{1}{\sum_{i=1}^{N_1} p_i} + \frac{1}{\sum_{i=1}^{N_1-1} p_i} + ... + \frac{1}{\sum_{i=1}^{N_1-k+1} p_i}. \tag{4.11}$$

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

$$\sum_{i=h}^{N_1} p_i \leq \sum_{j=h}^{N_1} p_{i_j} \leq \sum_{i=1}^{N_1-h+1} p_i, \ h = 1, 2..., N_1, \tag{4.12}$$

then

$$E(T_{k,N_1}|D_{1,2,...,k}) = \frac{1}{p} + \frac{1}{p - p_1} + \frac{1}{p - \sum_{i=1}^{2} p_i} + ... + \frac{1}{p - \sum_{i=1}^{k-1} p_i} \geq$$

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

$$\frac{1}{p} + \frac{1}{p - p_{N_1}} + \frac{1}{p - \sum_{i=N_1-1}^{N_1} p_i} + ... + \frac{1}{p - \sum_{i=N_1-k+2}^{N_1} p_i} =$$

$$E(T_{k,N_1}|D_{N_1,N_1-1,...,N_1-k+1})$$

and the proof is complete.

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

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

In other words, $E(T_{k,N_1}|D_{1,2,...,k})$ and $E(T_{k,N_1}|D_{N_1,N_1-1,...N_1-k+1})$ 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 $D_{i_1 i_2 ... i_k}$, $(i_1, i_2, ..., i_k) \in \Pi_k$, is $D_{1,2,...,k}$, i.e. the order of parasitism of k hosts mostly likely to occur is $1, 2, ..., k$.

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

The second part comes directly from the fact that

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

which follows from (4.2) and (4.12).

Propositions 4.2 and 4.3 prove that, if $p_1 \geq p_2 \geq \ldots \geq p_{N_1}$, then the most likely order of parasitization of k hosts in H_1 is the preferential order 1, 2, ..., k. Moreover the shortest 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 longest 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 $N_1 = N$. Recalling the definitions of these random variables and the random variables Y_n, we obtain the following relations

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

then

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

and

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

From (3.3), (3.5) and above equalities we see that

$$P(T_k > n) = \sum_{A \subset H, |A| \leq k-1} (-1)^{k-1-|A|} \binom{N-|A|-1}{k-1-|A|} p_A^n, \tag{4.13}$$

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

and

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

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

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

$E(T_k) = E(T_{k,N})$ is the expected number of draws required for k hosts are parasitized. Different expressions have been described for this expectation (Boneh and Hofri, 1989; Flajolet et al., 1992). From (4.13) it follows immediately that

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

$$\sum_{A\subset H,|A|\leq k-1}(-1)^{k-1-|A|}\binom{N-|A|-1}{k-1-|A|}\frac{1}{1-p_A}.$$

This expression was obtained in Flajolet et al. (1992). In Boneh and Hofri (1989) 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 (4.9)

$$E(T_N)=\sum_{(i_1 i_2 \ldots i_N)\in\Pi_N}\left(\sum_{r=0}^{N-1}\frac{1}{1-\sum_{j=1}^{r}p_{i_j}}\right)\frac{\prod_{i=1}^{N}p_i}{\prod_{k=1}^{N}\sum_{j=k}^{N}p_{i_j}} \tag{4.14}$$

where Π_N is the group of permutations of $\{1,2,\ldots N\}$. This expression for $E(T_N)$ is proposed in Boneh and Hofri (1989). 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 each.

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

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

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

$$E(Y_n) = N - \frac{(N-1)^n}{N^{n-1}},$$

$$Var(Y_n) = \frac{(N-1)(N-2)^n}{N^{n-1}} + \frac{(N-1)^n(N^{n-1} - (N-1)^n)}{N^{2n-3}}.$$

and the expected number of draws for k new hosts to be parasitized (4.9) is

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

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

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

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 $P(D_{i_1 i_2 ... i_k})$ are equal to $\frac{1}{N_1^{(k)}}$.

Two kinds of hosts. The two types of host situation is an idealization of the following cases. Hosts which are dead, either because they were previously parasitized or because they produced artifacts 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% (Casas, 1989; Casas et al, 2004). 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, ..., m has a probability α of being parasitized, and each hosts $m+1$, $m+2$, ..., N has a probability β of being parasitized, such that

$$\begin{aligned} p_1 = p_2 = ... = p_m = \alpha, \\ p_{m+1} = p_{m+2} = ... = p_N = \beta. \end{aligned} \tag{5.2}$$

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

$$P(S_{n1} = r_1, S_{n2} = r_2, ..., S_{nN} = r_N) = \frac{n!}{r_1! r_2! ... r_N!} \alpha^{\sum_{i \leq m} r_i} \beta^{\sum_{i > m} r_i}$$

$$0 \leq r_1 \leq n, 0 \leq r_2 \leq n, ..., 0 \leq r_N \leq n, \quad r_1 + r_2 + ... + r_N = n.$$

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

$$P(S_{ni_1} = r_{i_1}, S_{ni_2} = r_{i_2}, ..., S_{ni_h} = r_{i_h}) =$$

$$\frac{n!}{r_{i_1}!r_{i_2}!...r_{i_h}!(n - \sum r_{i_j})!}\, \alpha^{\sum_{i_j \leq m} r_{i_j}} \beta^{\sum_{i_j > m} r_{i_j}} \left(1 - \sum_{i_j \leq m} \alpha - \sum_{i_j > m} r_{i_j}\beta\right)^{n - \sum r_{i_j}}.$$

We will now calculate the expected number of parasitized hosts after n draws with this risk distribution, using the results obtained in Section 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, (3.6), 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 will name the hosts of the set H_1, hosts 1, 2, ..., N_1. Without any loss of generality, we can assume $p_1 = p_2 = ... = p_{m_1} = \alpha$ and $p_{m_1+1} = p_{m_1+2} = ... = p_{N_1} = \beta$. Let Π_k be the set of all k-permutations of the integers 1, 2, ..., N_1. For every $I = (i_1, i_2, ..., i_k) \in \Pi_k$, let $A_I \subset \{1, 2, ..., k\}$ be the set defined by $j \in A_I$ if $i_j \leq m_1$. It is clear that the probability $P(D_{i_1,i_2,...,i_k}) = P(D_I)$ given by (4.2) is, in this case,

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

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 $\overline{I}$ the equivalence class of I, and by $\overline{\Pi}_k$ the set whose elements are the equivalence classes of the elements of Π_k, that is

$$\overline{\Pi}_k = \left\{\overline{I} : I \in \Pi_k\right\}.$$

There are as many equivalence classes as subsets of $\{1,2,...,k\}$ 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

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

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

$$E(T_{k,N_1}) = \sum_{I\in\Pi_k}\left(\frac{1}{p}+\frac{1}{(p-\gamma_1)}+...+\frac{1}{(p-\sum_{j=1}^{k-1}\gamma_j)}\right) =$$

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

$$\sum_{\bar{I}\in\bar{\Pi}_k}\sum_{I\in\bar{I}}\left(\frac{1}{p}+\frac{1}{(p-\gamma_1)}+...+\frac{1}{(p-\sum_{j=1}^{k-1}\gamma_j)}\right)$$

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

$$\sum_{\bar{I}\in\bar{\Pi}_k} m_1^{(|A_I|)} n_1^{(k-|A_I|)}\left(\frac{1}{p}+\frac{1}{(p-\gamma_1)}+...+\frac{1}{(p-\sum_{j=1}^{k-1}\gamma_j)}\right)$$

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

$$\sum_{h=\max\{0,k-n_1\}}^{\min\{k,m_1\}}\sum_{|A_I|=h} m_1^{(h)} n_1^{(k-h)}\left(\frac{1}{p}+\frac{1}{(p-\gamma_1)}+...+\frac{1}{(p-\sum_{j=1}^{k-1}\gamma_j)}\right)$$

$$\frac{\alpha^{h}\beta^{k-h}}{p(p-\gamma_1)(p-\sum_{j=1}^{2}\gamma_j)...(p-\sum_{j=1}^{k-1}\gamma_j)}.$$

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

$$E(T_{k,N_1}|D_{1,2,...,k}) =$$

$$\frac{1}{m_1\alpha+n_1\beta}+\frac{1}{(m_1-1)\alpha+n_1\beta}+...+\frac{1}{(m_1-k+1)\alpha+n_1\beta},$$

if $k > m_1$, this upper bound is

$$E(T_{k,N_1}|D_{1,2,\ldots,k}) =$$

$$\frac{1}{m_1\alpha + n_1\beta} + \frac{1}{(m_1-1)\alpha + n_1\beta} + \ldots + \frac{1}{n_1\beta} +$$

$$\frac{1}{(n_1-1)\beta} + \ldots + \frac{1}{(n_1+m_1-k+1)\beta}.$$

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

$$E(T_{k,N_1}|D_{N_1,N_1-1,\ldots,N_1-k+1}) =$$

$$\frac{1}{m_1\alpha + n_1\beta} + \frac{1}{m_1\alpha + (n_1-1)\beta} + \ldots + \frac{1}{m_1\alpha + (n_1-k+1)\beta},$$

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

$$E(T_{k,N_1}|D_{N_1,N_1-1,\ldots,N_1-k+1}) =$$

$$\frac{1}{m_1\alpha + n_1\beta} + \frac{1}{m_1\alpha + (n_1-1)\beta} + \ldots + \frac{1}{m_1\alpha} +$$

$$\frac{1}{(m_1-1)\alpha} + \frac{1}{(n_1+m_1-k+1)\alpha}.$$

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

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

In the extreme case that there is only one host with a probability α of being parasitized and the others have probability β of being parasitized, we obtain the following expressions for $E(T_{k,N_1})$.

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

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

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

$$E(T_{k,N_1}) = (N_1-1)^{(k)}\left(\frac{1}{\alpha + (N_1-1)\beta} + \frac{1}{\alpha + (N_1-2)\beta} + \ldots + \frac{1}{\alpha + (N_1-k)\beta}\right)$$

$$\frac{\beta^k}{(\alpha+(N_1-1)\beta)(\alpha+(N_1-2)\beta)...(\alpha+(N_1-k)\beta)}+$$

$$(N_1-1)^{(k-1)}\sum_{j=1}^{k-1}\left(\frac{1}{\alpha+(N_1-1)\beta}+\frac{1}{\alpha+(N_1-2)\beta}+...+\frac{1}{\alpha+(N_1-j)\beta}+\right.$$

$$\left.\frac{1}{(N_1-j)\beta}+\frac{1}{(N_1-j-1)\beta}+...+\frac{1}{(N_1-k+1)\beta}\right)$$

$$\frac{\alpha\beta^{k-1}}{(\alpha+(N_1-1)\beta)(\alpha+(N_1-2)\beta)...(\alpha+(N_1-j)\beta)(N_1-j)\beta(N_1-j-1)\beta...(N_1-k+1)\beta}+$$

$$(N_1-1)^{(k-1)}\left(\frac{1}{\alpha+(N_1-1)\beta}+\frac{1}{\alpha+(N_1-2)\beta}+...+\frac{1}{\alpha+(N_1-k)\beta}\right)$$

$$\frac{\alpha\beta^{k-1}}{(\alpha+(N_1-1)\beta)(\alpha+(N_1-2)\beta)...(\alpha+(N_1-k)\beta)}.$$

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 20th century. A comprehensive review of the theory can be found in Marshall et al. (2011).

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 coupon collector problem seems to be largely unaware of it, but this result is not new and can be found in Wong and Yue (1973). This result constitutes the first part of Theorem 6.1. We give a proof more concise and clearer than previous proposal. Moreover, our method provides a precise result for strict Schur convexity. This refinement constitutes the second part of Theorem 6.1. 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 $p = (p_1, p_2, ..., p_N)$, 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 $(p_1, p_2, ..., p_N)$ in $\mathbb{R}^N$, we denote by $(p_{\bar{1}}, p_{\bar{2}}, ..., p_{\bar{N}})$ the N-uple obtained by permutation of p_i such that $p_{\bar{1}} \geq p_{\bar{2}} \geq \ldots \geq p_{\bar{N}}$.

The following definitions are given in Marshall et al. (2011).

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

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

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 $q = (q_1, q_2, ..., q_N) \in \mathbb{R}^N$. If $q_h < q_k$ 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 $q' = (q'_1, q'_2, ..., q'_N)$, 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 (Arnold, 1987). If we define $\alpha = 1 - \frac{\Delta}{q_k - q_h}$ 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 6.1. The following conditions are equivalent:

a) $p \prec q$,

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 in Muirhead (1902) for vectors of non-negative integer components.

Lemma 6.1. 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

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

and

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

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

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

and the lemma follows.

Lemma 6.2. 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 $m\ge 0$,

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

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

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

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

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

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

which proves the lemma.

Let $p=(p_1,p_2,...,p_N)$ 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 $0<h\le\frac{p_2-p_1}{2}$, $\alpha=1-\frac{h}{p_2-p_1}$. We then define a new risk distribution p' by applying a Robin Hood transfer as follows

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

We indeed have $p' \prec p$.

Theorem 6.1. 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 $N-1$,

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

which is equivalent to

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

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

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

which is equivalent to

$$P_p(T_{k+1} \leq n) < P_{p'}(T_{k+1} \leq n), \quad n = k+1, k+2, k+3... \tag{6.5}$$

where p' is defined by (6.1).

Proof. Let $H' = \{3, 4, ..., N\}$. According to (3.5) we have:

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

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

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

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

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

Then

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

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

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

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

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

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 6.1. 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 6.2 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 6.2. 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.

We can state the following corollaries.

Corollary 6.1. Let $p = (p_1, p_2, ..., p_N)$ and $q = (q_1, q_2, ..., q_N)$ be risk distributions on $H = \{1, 2, ..., N\}$. If $p \prec q$ then, for every $n \geq 1$ and every $k \geq 1$

$$P_p(Y_n \leq k) \leq P_q(Y_n \leq k), \tag{6.6}$$

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

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

- If $n \geq 2$, $k < n$ and $k < j$, then

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

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 6.1, 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 6.1. 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 $\mathcal{A} \subset \mathbf{R}^N$ 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 6.1 states that the map $p \to P_p(Y_n \leq k)$ is Schur-convex. This was

already proved in Wong and Yue (1973), and was stated as a conjecture in Anceaume et al. (2015).

Corollary 6.2. Let $u = (1/N, 1/N, ..., 1/N)$ be the uniform distribution on $H = \{1, 2, ..., N\}$ and $p = (p_1, p_2, ..., p_N)$ any other risk distribution on H. Then

$$P_u(Y_n \leq k) < P_q(Y_n \leq k), \quad k = 1, 2, ..., N-1, \quad n = k+1, k+2, ...$$

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

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

Remark 6.2. The results obtained in Corollary 6.1 and Corollary 6.2 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 $H = \{1, 2, ..., N\}$ is weakly stochastically dominated by the random variable Y_n defined on the probability space determined by q. Corollary 6.2 proves that the random variable Y_n defined on the probability space determined by the uniform distribution $u = (1/N, 1/N, ..., 1/N)$ 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 6.3. After the redaction of this section, we have seen a similar study in Anceaume et al. (2016). In particular, they prove inequalities (30) and (31) of Theorem 6.1. However, our contribution still presents a real interest, thanks to the quality of the argument based on use of fundamental formulas (6) and (7) in different contexts, and because we obtain cases of strict inequalities.

7 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.

The distribution functions of five variables Y_n are represented in graphic A of Figure 2. 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 $\{1, 2, ..., 12\}$ (so $N = 12$), p_1 is the uniform distribution, $p_i = (1/10i+2, ..., 1/10i+2, 10(i-1)+1/10i+2)$ for $i = 2$ and 3, and $p_i = (1/45(i-3)+12, ..., 1/45(i-3)+12, 1/45(i-3)+12, 45(i-3)+1/45(i-3)+12)$ 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 $k = 1, 2, ..., 11$, $i = 1, 2, 3, 4$.

The distribution functions of ten variables T_k are represented in every one of the graphics B and C in Figure 2. $N = 10$ and the risk distributions are the

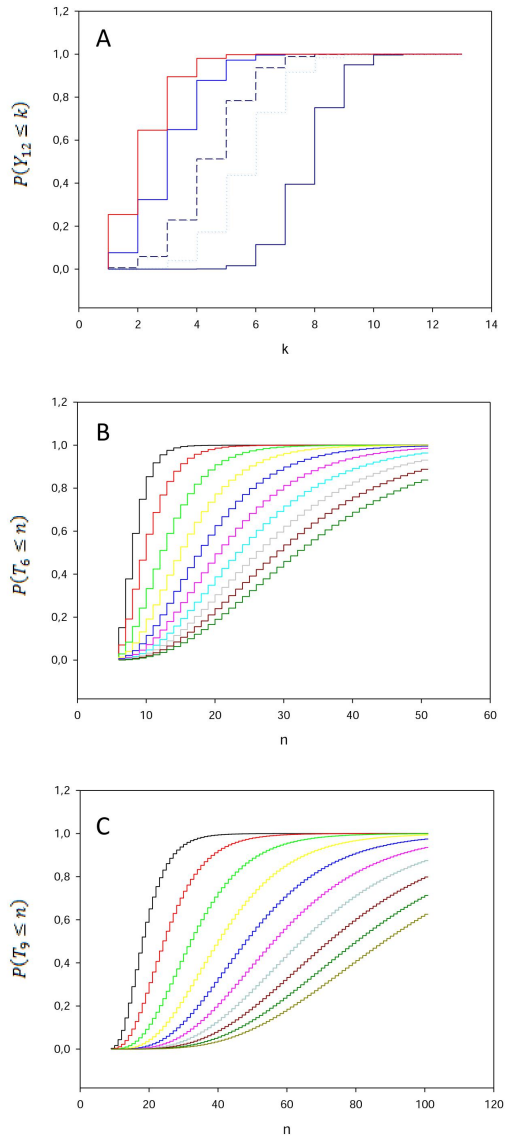


Figure 2. Graphic A: Distribution functions of five variables Y_{12} corresponding to five different risk distributions, p_1, p_2,..., p_5, satisfying $p_1 \prec p_2 \prec p_3 \prec p_4 \prec p_5$. It can be observed that $P_{p_i}(Y_{12} \leq k) < P_{p_{i+1}}(Y_{12} \leq k)$, for $k = 1, 2, ..., 11$, $i = 1, 2, 3, 4$. Graphic B: Distribution functions of ten variables T_6 corresponding to ten risk distributions, p_1, p_2, ..., p_{10} satisfying $p_1 \prec p_2 \prec ... \; p_9 \prec p_{10}$. Graphic C: Distribution functions of ten variables T_9 corresponding to the same previous risk distributions. In graphics B and C it can be observed that $P_{p_i}(T_k \leq n) > P_{p_{i+1}}(T_k \leq n)$, for $n = k, k+1, ... \; i = 1, 2, ...9$

same in both cases; p_1 is the uniform distribution and $p_i = (1/5(i+1), ..., 1/5(i+1), i/5(i+1), 4(i-1)+1/5(i+1))$, for $i = 2, 3, ..., 10$. For these risk distributions $p_1 \prec p_2 \prec ... \ p_9 \prec p_{10}$ is satisfied. In graphic B of Figure 2, $k = 6$ and the values of n lie between 6 to 50. In graphic C of Figure 2, $k = 9$ 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 $n = k, k+1, ...\ i = 1, 2, ...9$, in both graphics.

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, $p = (3/85, 3/85, 3/85, 3/85, 3/85, 12/85, 13/85, 13/85, 13/85, 19/85)$ and $q = (3/81, 4/81, 4/81, 4/81, 4/81, 5/81, 5/81, 5/81, 15/81, 32/81)$. In the first graphic $k = 5$, in the second $k = 8$ and in the third $k = 9$. In the last two cases the distribution functions cross. They do not cross in the first.

8 A conjecture on strong dominance

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

Definition 8.1. 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 Section 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 that one in which inequalities apply not only to the cumulative distribution functions, but also to the distributions themselves. We will 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 8.2. Let X and X' be two real random variables, defined on probabilities spaces (Ω, P) and (Ω', P'), respectively, and taking values in a denumerable set D. We say that the random variable X strongly stochastically

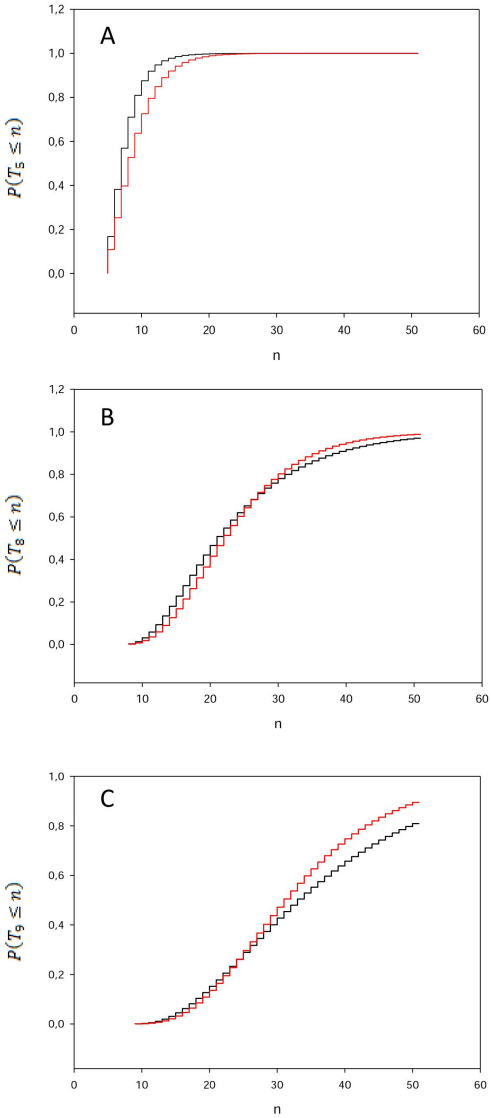


Figure 3. Comparison of distribution functions of random variables T_k corresponding to two unrelated risk distributions, i.e. neither $p \prec q$ nor $q \prec p$, to show how these distribution functions act in different ways depending on the value of k.

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 $P(X = d) \leq P'(X' = d)$,

•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 two risk distributions.

Acknowledgments

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 19320/IP/14. N.Z. is also grateful to the University François-Rabelais of Tours, for its support and hospitality.

References

Anceaume, E., Busnel, Y., & Sericola, B. (2015). New results on a generalized coupon collector problem using Markov chains. Journal of Applied Probability, 52(2), 405-418.

Anceaume, E., Busnel, Y., Shulte-Geers, E. & Sericola, B. (2016). Optimization results for a generalized coupon collector problem. Journal of Applied Probability, 53(2), 622-629.

Arnold, B. (1987). Majorization and the Lorenz Order: A Brief Introduction. Arnold, B. C.

Bailey L. L., MacKenzie, D. I., Nichols J. D. (2013). Special feature. Modelling demographic processes in marked populations: Proceedings of the Euring 2013 Analytical Meeting. Advances and applications of occupancy model. Methods in Ecology and Evolution, doi:10.111111/2041-210X.12100

Boneh, A., Hofri, M. (1989). The coupon-colector problem revisited. Computer Science Technical Report. Paper 807. http://docs-lib.purdue.edu/cstech/807.

Bunge, J., & Fitzpatrick, M. (1993). Estimating the number of species: a review. Journal of the American Statistical Association, 88(421), 364-373.

Casas, J. (1989). Foraging behaviour of a leafminer parasitoid in the field. Ecological Entomology, 14(3), 257-265.

Casas, J., Swarbrick, S., & Murdoch, W. W. (2004). Parasitoid behaviour: predicting field from laboratory. Ecological entomology, 29(6), 657-665.

Daley, D.J., Gani, J. & Gani, J.M. (2001). Epidemic modelling: an introduction. Cambridge University Press.

Dennehy, J. J. (2009). Bacteriophages as model organisms for virus emergence research. Trends in microbiology, 17(10), 450-457.

Dixon, C. J. (2006). A means of estimating the completeness of haplotype sampling using the Stirling probability distribution. Molecular Ecology Notes, 6(3), 650-652.

Donnelly, P. (1986). Partition structures, Polya urns, the Ewens sampling formula, and the ages of alleles. Theoretical Population Biology, 30(2), 271-288.

Doumas, A.V. (2015). How many trials does it take to collect all different types of a population with probability p?. Journal of Applied Mathematics and Bioinformatics, 5(3), 1-14.

Ewens, W. J. (1972). The sampling theory of selectively neutral alleles. Theoretical population biology, 3(1), 87-112.

Feller, V. (1968). An Introduction to Probability Theory and Its Applications: Volume One. John Wiley & Sons.

Fiske, W. F. (1910). Superparasitism: an important factor in the natural control of insects. Journal of Economic Entomology, 3(1), 88-97.

Flajolet, P., Gardy, D., & Thimonier, L. (1992). Birthday paradox, coupon collectors, caching algorithms and self-organizing search. Discrete Applied Mathematics, 39(3), 207-229.

Hassell, M. (2000). The spatial and temporal dynamics of host-parasitoid interactions. Oxford University Press.

Hemerik, L., Van der Hoeven, N., & van Alphen, J. J. (2002). Egg distributions and the information a solitary parasitoid has and uses for its oviposition decisions. Acta biotheoretica, 50(3), 167-188.

Hernandez-Suarez, C. M., & Hiebeler, D. (2012). Modeling species dispersal with occupancy urn models. Theoretical Ecology, 5(4), 555-565.

Huillet, T., & Paroissin, C. (2009). Sampling from Dirichlet partitions: estimating the number of species. Environmetrics, 20(7), 853-876.

Ives, A.R., Schooler, S.S., Jagar, V.J., Grbic, M. &Settle, W.H. (1999). Variability and parasitoid foraging efficiency: a case study of pea aphids and Aphidius ervi. The American Naturalist, 154(6), 652-673.

Kershenbaum A, Freeberg TM, Gammon DE. 2015. Estimating vocal repertoire size is like collecting coupons: A theoretical framework with heterogeneity in signal abundance. Journal of Theoretical Biology, 373(2015): 1-11.

Keeling, M.J., & Rohani, P. (2008). Modelling infectious diseases in humans and animals. Ptinceton University Press.

Lloyd-Smith, J.O., Schreiber, S.J., Kopp, P.E. & Getz, W.M. (2005). Superspreading and the effect of individual variation on disease emergence. Nature, 438(7066), 355-359.

Marshall, A. W.,Olkin, I., Arnold, B. C. (2011) Inequalities: Theory of majorization and its applications. Second edition. Springer Series in Statistics. Springer, New York.

May, R.M. (1978). Host-parasitoid systems in patchy environments: a phenomenological model. The Journal of Animal Ecology, 833-844.

MacArthur, R. H. (1957). On the relative abundance of bird species. Proceedings of the National Academy of Sciences of the United States of America, 43(3), 293-295.

Montovan, K.J., Couchoux, C., Jones, L.E., Reeve, H.K., van Nouhuys, S. (2015) The puzzle of partial resource use by a parasitoid wasp. The American Naturalist, 185(4) 538-550.

Muirhead, R. F. (1902). Some methods applicable to identities and inequalities of symmetric algebraic functions of n letters. Proceedings of the Edinburgh Mathematical Society, 21, 144-162.

Murdoch, W.W., Briggs, C.J. & Nisbet, R.M. (2013). Consumer-Resource Dynamics (MPB-36). Princeton University Press.

Neal, P., & Moriary, J. (2009). Sampling Efficiency and Biodiversity.

Simpson, E. H. (1949). Measurement of diversity. Nature, 163, 688.

Singh, A., Murdoch, W.W. & Nisbet, R.M. (2009). Skewed attacks, stability, and host suppression. Ecology, 90(6), 1679-1686.

Tenaillon, O., Rodríguez-Verdugo, A., Gaut, R. L., McDonald, P., Bennett, A. F., Long, A. D., & Gaut, B. S. (2012). The molecular diversity of adaptive convergence. Science, 335(6067), 457-461.

Thompson, W. R. (1924). La théorie mathématique de l'action des parasites entomophages et le facteur du hasard. Ann. Fac. Sci. Marseille, 2(2), 69-89.

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. Nature communications, 6.

Wong, C. K., & Yue, P. C. (1973). A majorization theorem for the number of distinct outcomes in N independent trials. Discrete Mathematics, 6(4), 391-398.