Sampling from Dirichlet populations: estimating the number of species

Thierry Huillet, Christian Paroissin

Sampling from Dirichlet partitions: estimating the number of species

Thierry Huillet[1] and Christian Paroissin[2]

[1] *Université de Cergy-Pontoise, Laboratoire de Physique Théorique et Modélisation - UMR CNRS 8089, 2 avenue Adolphe Chauvin, 95032 Cergy-Pontoise, France. Email:* huillet@ptm.u-cergy.fr
[2] *Université de Pau et des Pays de l'Adour, Laboratoire de Mathématiques Appliquées - UMR CNRS 5142, Avenue de l'Université, 64013 Pau cedex, France. Email:* cparoiss@univ-pau.fr

Last version: September 24, 2008

Consider the random Dirichlet partition of the interval into n fragments with parameter $\theta > 0$. We recall the unordered Ewens sampling formulae from finite Dirichlet partitions. As this is a key variable for estimation purposes, focus is on the number of distinct visited species in the sampling process. These are illustrated in specific cases. We use these preliminary statistical results on frequencies distribution to address the following sampling problem: what is the estimated number of species when sampling is from Dirichlet populations? The obtained results are in accordance with the ones found in sampling theory from random proportions with Poisson-Dirichlet distribution. To conclude with, we apply the different estimators suggested to two different sets of real data.

Keywords: random discrete distribution, Dirichlet partition, GEM, Ewens sampling formulae, estimated number of species from sampling
AMS 2000 Classification: 60G57, 62E17, 60K99, 62E15, 62E20

1 Introduction

Dirichlet partition of an interval can be viewed as a generalization of some classical models in ecological statistics. For example, on the one hand, when $\theta = 1$, the Dirichlet partition corresponds to the broken-stick model (see Feller (1966), pages 22-24), one of the most famous stochastic model of relative species abundance studied by McArthur (1957) (see also Tokeshi (1993) for an exhaustive survey on species abundance models). On the other hand, when θ goes to infinity, the Dirichlet partition is deterministic and uniform and when θ goes to 0, jointly with the numbers of fragments going to infinity, the ordered version of Dirichlet partition identifies with the Poisson-Dirichlet (PD) partition and corresponds to the Fisher's log-series model. These relationships between all models cited above was already pointed out early by Simpson (1949) (but the term "Dirichlet distribution" was coined by Wilks (1962) many years later).

The organization of this manuscript is the following. In Section 2 we recall the Ewens sampling formulae when sampling is from finite Dirichlet partitions. Consider the random Dirichlet partition of the interval into n fragments with parameter $\theta > 0$. Elementary properties of its $D_n(\theta)$ distribution are first recalled in section 2.1. Section 2.2 describes some motivational sampling problems from Dirichlet proportions. Some generalities about sampling from Dirichlet partition are first proved in subsection 2.3. Subsection 2.4 is devoted to the Ewens sampling formulae when sampling is from Dirichlet partition $D_n(\theta)$. Here the order in which sequentially sampled species arise is irrelevant. Similarly, the second Ewens sampling formula under the same hypothesis (as a problem of random partitioning of the integers). As corollaries to these results, assuming $n \uparrow \infty$, $\theta \downarrow 0$ while $n\theta = \gamma > 0$, the usual well-known sampling formulae will be deduced in each case when sampling is from $PD(\gamma)$ distribution. These general sampling formulae are also illustrated in detail in two particular cases: the Bose-Einstein case (when $\theta = 1$) and the Maxwell-Bolztmann case (when θ tends to infinity). As this is the key variable for estimation purposes, focus is also made on the number of distinct visited species in the sampling process, in each case.

Section 3 concerns the statistical problem of estimating the number of distinct species in a Dirichlet population. A maximum-likelihood estimator is developed, which is derived from sampling formulae recalled in the previous section. We recall also the minimum variance one suggested by Keener *et al.* (1987). For some particular classes of Dirichlet partitions, we supply simpler expressions for these estimators. We study some related statistical questions like stopping rule in the sampling process and goodness of fit. In the last subsection, we explore the difficult

problem of estimating jointly n and θ.

At least section 4 is devoted to applications to real data. The first two data sets concern word usage by two authors (Keener *et al.*, 1987) while the second two data sets deal with tropical beetles species (Janzen, 1973).

2 Sampling from Dirichlet proportions

First we recall basic properties about the symmetric Dirichlet distributions. Second we give some motivation about sampling with this distribution. Then we recall sampling formulae that will be useful later.

2.1 Dirichlet partition of the interval

Consider the following random partition into n fragments of the unit interval. Let $\theta > 0$ be some parameter and assume that the random fragment sizes $\mathbf{S}_n = (S_1, \ldots, S_n)$ (with $\sum_{m=1}^n S_m = 1$) are exchangeable and distributed according to the (symmetric) Dirichlet $D_n(\theta)$ p.d.f. which is defined on the simplex, i.e.

$$f_{S_1,\ldots,S_n}(s_1, \ldots, s_n) = \frac{\Gamma(n\theta)}{\Gamma(\theta)^n} \prod_{m=1}^n s_m^{\theta-1} \delta_{\left(\sum_{m=1}^n s_m - 1\right)} \, . \tag{1}$$

Alternatively the distribution of $\mathbf{S}_n = (S_1, \ldots, S_n)$ can also be characterized by its joint moment function

$$\forall (q_1, \ldots, q_n) \in \mathbb{R}^n \, , \quad \mathbf{E}_\theta \left[\prod_{m=1}^n S_m^{q_m} \right] = \frac{\Gamma(n\theta)}{\Gamma(n\theta + \sum_{m=1}^n q_m)} \prod_{m=1}^n \frac{\Gamma(\theta + q_m)}{\Gamma(\theta)} \, .$$

We shall put $\mathbf{S}_n \overset{d}{\sim} D_n(\theta)$ if $\mathbf{S}_n$ is Dirichlet distributed with parameter θ. In such case $S_m \overset{d}{=} S_n$, for any $m \in \{1, \ldots, n\}$, independently of m and the individual fragment sizes are all identically distributed. Their common p.d.f. on the interval $(0, 1)$ is the beta distribution with parameter $(\theta, (n-1)\theta)$. As a result, parameter θ interprets as a "precision" parameter indicating how concentrated the distribution of $\mathbf{S}_n$ is around its mean $(\frac{1}{n}, \ldots, \frac{1}{n})$: the larger θ is, the more the distribution of $\mathbf{S}_n$ is concentrated around its mean. Indeed as one can check that, for any $m \in \{1, \ldots, n\}$, $\mathbf{E}(S_m) = \frac{1}{n}$ and $\text{var}(S_m) = \frac{n-1}{n^2(n\theta+1)}$.

In the random division of the interval as in equation (1), although all fragments are identically distributed with expected sizes of order $\frac{1}{n}$, the smallest fragment size grows like $n^{-(\theta+1)/\theta}$ while the size of the largest is of order $\frac{1}{n\theta} \log(n \log^{\theta-1} n)$. Consistently, the smaller θ is, the larger (resp. the smaller) the largest (resp. the smallest) fragment size is: hence, the smaller θ is, the more the values of the S_m are disparate with high probability. Let $\mathbf{S}_{(n)} = (S_{(1)}, \ldots, S_{(n)})$ be the ordered version of $\mathbf{S}_n$ with $S_{(1)} > \cdots > S_{(n)}$. The smaller the parameter θ is, the more the size of the largest fragment $S_{(1)}$ tends to dominate the other ones. On the contrary, for large values of θ, the fragment sizes look more homogeneous and distribution equation (1) concentrates on its centre $(\frac{1}{n}, \ldots, \frac{1}{n})$. For large θ, the diversity of the partition is small.

When $\theta = 1$, the partition corresponds to the standard uniform random partition model of the interval. When $\theta \uparrow \infty$, $\mathbf{S}_n$ approaches the deterministic partition of the interval into n equal parts with sizes $1/n$. Although $\mathbf{S}_n$ has a degenerate weak limit, when $n \uparrow \infty$, $\theta \downarrow 0$ while $n\theta = \gamma > 0$, this situation is worth being considered. Indeed, many interesting statistical features emerge from the fact that in such asymptotic regime $\mathbf{S}_{(n)}$ converges to $\mathbf{S}_{(\infty)}$ having the Poisson-Dirichlet distribution $PD(\gamma)$ with parameter γ (see Kingman, 1975). These three situations will be referred later respectively by Bose-Einstein, Maxwell-Boltzmann and Kingman cases.

2.2 Sampling: motivations

We shall be interested in sampling problems from random partition $\mathbf{S}_n$, where $\mathbf{S}_n \overset{d}{\sim} D_n(\theta)$. Since $\mathbf{S}_n$ is random, sampling occurs in a random environment. Dirichlet distributions are ubiquitous in the natural sciences and this is why we chose this model for the random probabilities $\mathbf{S}_n$. We refer to Vlad *et al.* (2001) where it is shown that Dirichlet distributions may be seen as limit laws of certain "dilution" processes, and also that they maximize entropy under constraints, satisfy some scale-invariance property, etc. Due to its specific statistical properties as a random partition, many combinatorial issues arising in this sampling context can receive a proper and exact analytical answer. We shall illustrate this point.

Sampling from $\mathbf{S}_n$ consists in a recursive k throw of iid uniformly distributed random variables on the interval. It is said that fragment number m is visited by some uniform throw if its hits the interval $[S_1 + \cdots + S_{m-1}, S_1 + \cdots + S_m]$ of length S_m. Before giving some technical details of the sampling problem, let us list some motivating concrete images of the sampling problem from $\mathbf{S}_n$:

- S_m could be the random abundance of species m from a population with n "animals". Some sampling process starts when a biologist records each new species met at each of his k measurement campaigns.
- S_m could be the random size of district number m of some city, with $m \in \{1, \ldots, n\}$ (e.g. with $n = 20$ for Paris). An unfriendly sampling process could be a scattered shot bombing with k bombs.
- S_m could be the random popularity of book m in a library with n books. The sampling process is when k consecutive readers borrow books from the library while respecting their popularities.
- S_m could be the random probability to be born on day m, with $n = 365$. A classroom with k students is a $k-$sample from $\mathbf{S}_n$.

We now give a non-exhaustive list of statistical problems of interest in this context:

- *Abundance estimation:* given a sample size k, estimate the number n of species and/or the parameter θ, exploiting for example the information on the empirical number p of distinct visited species or from the knowledge of the empirical probability to visit twice the same species.
- *Match box problem:* what is the state of fragment occupancies if sequential sampling process is stopped when some fragment has received c visits for the first time (if $n = 2$, this is the randomized Banach match box problem). In particular what is the probability that some cell is empty at this stopping time?
- *Birthday problem:* what is the sample size until the first visit to two species of the same type?
- *Coupon collector problem:* what is the sample size until all species have been visited at least once (or r times)?
- *Law of succession:* given the random number of occurrences of species m in the $k-$sample and the number p of distinct visited species, what is the probability to discover a new species in a $k+1$ sample? or what is the probability that the $(k+1)$-th sample is one from the previously encountered species already met a certain amount of times.

We shall now be more precise and treat rigorously some of the raised problems, starting with the sampling problem before focusing on the occupancy distributions. In the next section, using these results, we shall come to the important problem of estimating n when it is unknown.

2.3 Sampling: preliminaries and generalities

Let $(U_1, \ldots, U_k)$ be k iid uniform throws on $\mathbf{S}_n$. Let

$$\mathbf{K}_{n,k} = \left(\mathcal{K}_{n,k}(1), \ldots, \mathcal{K}_{n,k}(n)\right) \geqslant 0$$

be an integral-valued random vector which counts the number of visits to the different fragments in a k-sample. Hence, if M_l is the random fragment number (or label) in which the l-th trial falls, $\mathcal{K}_{n,k}(m) = \sum_{l=1}^{k} \mathbf{I}(M_l = m)$, $m \in \{1, \ldots, n\}$. Under our assumptions, for instance, we have that $\mathbf{P}_\theta(M_l = m \mid \mathbf{S}_n) = S_m$ (the random probability to visit species m is equal to its abundance) but also that the conditional probability to observe species number m in a k-sample is

$$\pi_m(S_m) = \mathbf{P}_\theta\left(\mathcal{K}_{n,k}(m) > 0 \mid \mathbf{S}_n\right) = 1 - (1 - S_m)^k. \tag{2}$$

Let us now focus our attention on the distribution of the occupancies $\mathbf{K}_{n,k}$. With $\sum_{m=1}^{n} k_m = k$ and $\mathbf{k}_n = (k_1, \ldots, k_n) \geqslant 0$, $\mathbf{K}_{n,k}$ follows the conditional multinomial distribution:

$$\mathbf{P}_\theta\left(\mathbf{K}_{n,k} = \mathbf{k}_n \mid \mathbf{S}_n\right) = \frac{k!}{\prod_{m=1}^{n} k_m!} \prod_{m=1}^{n} S_m^{k_m}.$$

Averaging over $\mathbf{S}_n$, using Dirichlet integrals, one finds

$$\mathbf{P}_\theta\left(\mathbf{K}_{n,k} = \mathbf{k}_n\right) = \mathbf{E}\mathbf{P}_\theta\left(\mathbf{K}_{n,k} = \mathbf{k}_n \mid \mathbf{S}_n\right) = \frac{\prod_{m=1}^{n} [\theta]_{k_m}^{n}}{[n\theta]_k},$$

where $[\theta]_k = (\theta)_k/k!$ and $(\theta)_k = \theta(\theta+1)\cdots(\theta+k-1)$,for any $k \geqslant 1$, with $(\theta)_0 = 1$. Applying Bayes formula, the posterior distribution of $\mathbf{S}_n$ given $\mathbf{K}_{n,k} = \mathbf{k}_n$ is determined by its p.d.f. at point $\mathbf{s}_n$on the simplex as

$$f_{\mathbf{S}_n}(\mathbf{s}_n \mid \mathbf{K}_{n,k} = \mathbf{k}_n) = \frac{\Gamma(n\theta+k)}{\prod_{m=1}^n \Gamma(\theta+k_m)} \prod_{m=1}^n s_m^{(\theta+k_m)-1}\, \delta_{(\sum_{m=1}^n s_m - 1)}.$$

This shows, as it is well-known, that

$$\mathbf{S}_n \mid \mathbf{K}_{n,k} = \mathbf{k}_n \overset{d}{\sim} D_n\left(\theta\mathbf{1} + \mathbf{k}_n\right),$$

an asymmetric Dirichlet distribution with parameters $\theta\mathbf{1} + \mathbf{k}_n = (\theta + k_1, \ldots, \theta + k_n)$. Furthermore,

$$\forall m \in \{1, \ldots, n\}\,, \quad \mathbf{E}_\theta\left(S_m \mid \mathbf{K}_{n,k} = \mathbf{k}_n\right) = \frac{\theta + k_m}{n\theta + k}\,.$$

This suggests a recursive approach to the sampling formula where successive sample are drawn from the corresponding iterative posterior distributions. More specifically, let $(M_1, \ldots, M_k) \in \{1, \ldots, n\}^k$ be the numbers of the successive fragments thus drawn. Then,

$$\mathbf{P}_\theta\left(M_1 = m_1\right) = \mathbf{E}\left(\mathbf{P}_\theta\left(M_1 = m_1\right) \mid \mathbf{S}_n\right) = \mathbf{E}_\theta\left(S_{m_1}\right) = \frac{\theta}{n\theta} = \frac{1}{n}\,,$$

$$\mathbf{P}_\theta\left(M_2 = m_2 \mid M_1\right) = \frac{\theta + \mathbf{I}\left(M_1 = m_2\right)}{n\theta + 1}$$

and

$$\mathbf{P}_\theta\left(M_k = m_k \mid M_1, \ldots, M_{k-1}\right) = \frac{\theta + \sum_{l=1}^{k-1} \mathbf{I}\left(M_l = m_k\right)}{n\theta + k - 1}.$$

Proceeding in this way, the joint distribution of $(M_1, \ldots, M_k)$ reads

$$\mathbf{P}_\theta\left(M_1 = m_1, \ldots, M_k = m_k\right) = \frac{\theta}{n\theta} \prod_{l=1}^{k-1} \frac{\theta + \sum_{j=1}^{l} \mathbf{I}\left(M_j = m_{l+1}\right)}{n\theta + l} = \frac{\prod_{m=1}^n (\theta)_{k_m}}{(n\theta)\, k},$$

where $k_m = \sum_{l=1}^k \mathbf{I}\left(m_l = m\right)$. This distribution being invariant under permutations of the entries, the sequence $(M_1, \ldots, M_k)$ is exchangeable. It is called a Pòlya urn sequence. We now prove the following convergence result:

Lemma 2.1 *Almost surely and in distribution, the following convergence holds:*

$$\frac{\mathbf{K}_{n,k}}{k} \xrightarrow[k\to\infty]{a.s.} \mathbf{S}_n\,.$$

Proof Let us first prove the convergence in distribution. The joint conditional generating function of $\mathbf{K}_{n,k}$ reads

$$\mathbf{E}_\theta\left(\prod_{m=1}^n u_m^{\mathcal{K}_{n,k}(m)} \mid \mathbf{S}_n\right) = \left(\sum_{m=1}^n u_m S_m\right)^k,$$

which is homogeneous with degree k allowing to compute $\mathbf{E}_\theta\left(\prod_{m=1}^n u_m^{\mathcal{K}_{n,k}(m)}\right)$. Further, defining $\widetilde{X}_m = X_m / \sum_{m=1}^n X_m$, where $X_m \overset{d}{\sim}$ gamma(θ), for all $m \in \{1, \ldots, n\}$, using independence between $(\widetilde{X}_1, \ldots, \widetilde{X}_n)$ and $\sum_{m=1}^n X_m$ and recalling that $(\widetilde{X}_1, \ldots, \widetilde{X}_n)$ has the Dirichlet distribution $D_n(\theta)$, we get

$$\begin{aligned}
\mathbf{E}_\theta\left(\prod_{m=1}^n u_m^{\mathcal{K}_{n,k}(m)/k}\right) &= \frac{\Gamma(n\theta)}{\Gamma(n\theta+k)} \mathbf{E}_\theta\left[\left(\sum_{m=1}^n u_m^{1/k} X_m\right)^k\right] \\
&\underset{k\to\infty}{\sim} \frac{\Gamma(n\theta)}{\Gamma(n\theta+k)} \mathbf{E}_\theta\left[\left(\sum_{m=1}^n X_m\right)^k \left(1 + \frac{1}{k}\sum_{m=1}^n \widetilde{X}_m \log u_m\right)^k\right] \\
&\underset{k\to\infty}{\sim} \mathbf{E}_\theta\left(\prod_{m=1}^n u_m^{\widetilde{X}_m}\right) = \mathbf{E}_\theta\left(\prod_{m=1}^n u_m^{S_m}\right).
\end{aligned}$$

Thus,

$$\frac{\mathbf{K}_{n,k}}{k} \xrightarrow[k\to\infty]{a.s.} \mathbf{S}_n\,.$$

By applying the strong law of large numbers (conditionally given $\mathbf{S}_n$), the above convergence in distribution also holds almost surely. This shows that $\mathbf{K}_{n,k}/k$ can be used as an consistent estimator of $\mathbf{S}_n$. □

2.4 Ewens sampling formulae for Dirichlet partitions

Ewens Sampling Formula (ESF) gives the distribution of alleles (different types of genes) in a sample with size k from the Poisson-Dirichlet partitioning $PD(\gamma)$. Alternatively, it can be described in terms of sequential sampling of animals from a countable collection of distinguishable species drawn from $PD(\gamma)$. It provides the probability of the partition of a sample of, say k, selectively equivalent genes into a number of alleles as population size becomes indefinitely large. When the order of appearance of sequentially sampled species does not matter, we are led to the first ESF for unordered sequences. A second equivalent way to describe the sample is to record the number of species in the k-sample with exactly i representatives, for $i \in \{0, \ldots, k\}$. When doing this while assuming the species have random frequencies following $PD(\gamma)$ distribution, we are led to a second Ewens Sampling Formula.

We recall here the exact expressions of both first and second Ewens sampling formulae, when sampling is first from finite Dirichlet random partitions with n fragments. Here, the order in which the consecutive animals are being discovered in the sampling process is irrelevant. In the sampling formulae, the joint event that there are p distinct fragments visited will also be taken into account. These sampling formulae give both ESF formulae from $PD(\gamma)$ when passing to the Kingman limit.

Let $\mathbf{S}_n$ be the above Dirichlet random partition with parameter $\theta > 0$. Let $k > 1$ and $(U_1, \ldots, U_k)$ be k iid uniform random throws on $[0, 1]$. Let then $(M_1, \ldots, M_k)$ be the (conditionally iid) corresponding animals species, with common conditional and unconditional distributions:

$$\forall m \in \{1, \ldots, n\}\,, \qquad \mathbf{P}_\theta\left(M = m \mid \mathbf{S}_n\right) = S_m\,,$$

and

$$\forall m \in \{1, \ldots, n\}\,, \qquad \mathbf{P}_\theta\left(M = m\right) = \mathbf{E}\left[\mathbf{P}_\theta\left(M = m \mid \mathbf{S}_n\right)\right] = \mathbf{E}_\theta\left(S_m\right) = \frac{1}{n}\,.$$

Recall $\mathcal{K}_{n,k}(m) = \sum_{l=1}^{k} \mathbf{I}\left(M_l = m\right)$ counts the random number of occurrences of species m in the k-sample and let $P_{n,k} = \sum_{m=1}^{n} \mathbf{I}\left(\mathcal{K}_{n,k}(m) > 0\right)$ count the number of distinct species which have been visited in the k-sampling process.

There are two occupancies variables of interest: the first one will lead to the first Ewens sampling formula while the second corresponds to the second Ewens sampling formula.

I. For any $q \in \{1, \ldots, p\}$, $\mathcal{B}_{n,k}(q) > 0$ is the numbers of animals of species q where the $P_{n,k} = p$ species observed were labelled in an arbitrary way (independently of the sampling mechanism). Thus $\mathcal{B}_{n,k}$ differs from $\mathcal{K}_{n,k}$ in the sense that all the components of $\mathcal{B}_{n,k}$ are positive.

II. For any $i \in \{0, \ldots, k\}$, $\mathcal{A}_{n,k}(i)$ is the number of species in the k-sample with i representatives, i.e.

$$\mathcal{A}_{n,k}(i) = \#\{m \in \{1, \ldots, n\} : \mathcal{K}_{n,k}(m) = i\} = \sum_{m=1}^{n} \mathbf{I}\left(\mathcal{K}_{n,k}(m) = i\right).$$

Then $\sum_{i=0}^{k} \mathcal{A}_{n,k}(i) = n$ is the (unknown) number of fragments and $\sum_{i=1}^{k} \mathcal{A}_{n,k}(i) = p$ is the number of fragments visited by the k-sample and $\mathcal{A}_{n,k}(0)$ the number of unvisited ones. Note that $\sum_{i=1}^{k} i\mathcal{A}_{n,k}(i) = k$ is the sample size. The random vector $\left(\mathcal{A}_{n,k}(1), \ldots, \mathcal{A}_{n,k}(k)\right)$ is called the fragment vector count or the species vector count in biology, see Ewens (1990).

For each of the two sampling problems, we easily obtain the Ewens sampling formulae from finite partitions $\mathbf{S}_n$ drawn from Dirichlet distribution. The following result can be found in Huillet (2005) (see also Ewens (1972) for the PD case)

Theorem 2.1

I *For any* $(b_1, \ldots, b_p)$ *such that* $\forall q \in \{1, \ldots, p\}$, $b_q \geqslant 1$ *and* $\sum_{q=1}^{p} b_q = k$, *we have*

$$\mathbf{P}_\theta\left(\mathcal{B}_{n,k}(1) = b_1, \ldots, \mathcal{B}_{n,k}(p) = b_p; P_{n,k} = p\right) = \binom{n}{p} \frac{k!}{\prod_{q=1}^{p} b_q!} \frac{1}{(n\theta)_k} \prod_{q=1}^{p} (\theta)_{b_q} \tag{3}$$

II *For any* $(a_1, \ldots, a_k) \geqslant 0$ *such that* $\sum_{i=1}^{k} i a_i = k$ *and* $\sum_{i=1}^{k} a_i = p$, *we have*

$$\mathbf{P}_\theta\left(\mathcal{A}_{n,k}(1) = a_1, \ldots, \mathcal{A}_{n,k}(k) = a_k; P_{n,k} = p\right) = \frac{n!}{(n-p)!} \frac{k!}{\prod_{i=1}^{k} \left(i!^{a_i} a_i!\right)} \frac{1}{(n\theta)_k} \prod_{i=1}^{k} (\theta)_i^{a_i} \tag{4}$$

From equation (3) or equivalently from equation (4), one can obtain the marginal distribution of $P_{n,k}$:

Theorem 2.2 *For any $p \geqslant 1$,*

$$\mathbf{P}_\theta\left(P_{n,k}=p\right)=\frac{n!}{(n-p)!}\frac{1}{(n\theta)_k}B_{k,p}\left(\theta\right) \tag{5}$$

where

$$B_{k,p}(\theta)=\frac{k!}{p!}\sum_{\substack{b_q\geqslant 1\\ \sum_{q=1}^p b_q=k}}\prod_{q=1}^{p}\frac{(\theta)_{b_q}}{b_q!}=\sum_{\substack{a_i\geqslant 0\\ \sum_{i=1}^k a_i=p\\ \sum_{i=1}^k i a_i=k}}\frac{k!}{\prod_{i=1}^k(i!^{a_i}a_i!)}\prod_{i=1}^{k}(\theta)_i^{a_i}\ .$$

We recall below a straightforward representation of the probability $\mathbf{P}_\theta(P_{n,k}=p)$ under the form of an alternate sum (see for example Keener *et al.*, pages 1471–1472).

Proposition 2.1 *For any $m\in\{0,\ldots,n-1\}$, let $\langle\theta\rangle_{n,k;m}=\frac{((n-m)\theta)_k}{(n\theta)_k}$. The distribution of $P_{n,k}$ is given by*

$$\mathbf{P}_\theta\left(P_{n,k}=p\right)=\sum_{q=1}^{p}(-1)^{p-q}\binom{n}{p}\binom{p}{q}\langle\theta\rangle_{n,k;n-q}\,. \tag{6}$$

Let us now focus on two problems related to sampling as explained previously.

The law of succession We would like to briefly recall a related question raised in Donnelly (1986) and Ewens (1996), concerning the law of succession.

1. Let the "M_{k+1} is new" denote the event that M_{k+1} is none of the previously observed species. One can prove that
$$\mathbf{P}_\theta(M_{k+1}\text{is new}\mid\mathcal{B}_{n,k}(1)=b_1,\ldots,\mathcal{B}_{n,k}(p)=b_p;P_{n,k}=p)=\frac{(n-p)\theta}{n\theta+k}, \tag{7}$$
which is independent of cell occupancies $b_1,\ldots,b_p$ but depends on the number p of distinct species already visited by the k-sample. With $k=p=2$, this is the probability that the first two random throws will visit any two distinct species. The complementary probability that it does not is thus $1-\frac{(n-1)\theta}{n\theta+1}=\frac{\theta+1}{n\theta+1}$. The probability to visit any fragment twice varies between 1 and $\frac{1}{n}$ when θ varies from 0 (the largest fragment dominates) to infinity (fragment sizes distribution approaches $\frac{1}{n}$).

2. Similarly, let the event "M_{k+1} is a species seen b_r times" denote the fact that the $(k+1)$-th sample is one from the previously encountered fragment already visited b_r times. We easily get
$$\mathbf{P}_\theta(M_{k+1}\text{ is a species seen }b_r\text{times}\mid\mathcal{B}_{n,k}(1)=b_1,\ldots,\mathcal{B}_{n,k}(p)=b_p;P_{n,k}=p)=\frac{\theta+b_r}{n\theta+k} \tag{8}$$
which is as previously independent of cell occupancies but also of the number p of distinct species.

The number of distinct observations From equations (7) and (8), we also have the transition probabilities

$$\mathbf{P}_\theta(P_{n,k+1}=p+1\mid P_{n,k}=p)=\frac{(n-p)\theta}{n\theta+k}$$

and

$$\mathbf{P}_\theta(P_{n,k+1}=p\mid P_{n,k}=p)=\frac{\sum_{r=1}^{p}(\theta+b_r)}{n\theta+k}=\frac{p\theta+k}{n\theta+k}.$$

It follows that we have the following recursion for the distribution of $P_{n,k}$:

$$\mathbf{P}_\theta(P_{n,k+1}=p)=\frac{(n-p+1)\theta}{n\theta+k}\mathbf{P}_\theta(P_{n,k}=p-1)+\frac{p\theta+k}{n\theta+k}\mathbf{P}_\theta(P_{n,k}=p)\ .$$

Using equation (5), we obtain the following triangular recurrence for the quantities $B_{k,p}(\theta)$

$$B_{k+1,p}\left(\theta\right)=\theta B_{k,p-1}\left(\theta\right)+\left(p\theta+k\right)B_{k,p}\left(\theta\right)\ .$$

These should be considered with boundary conditions

$$B_{k,0}(\theta) = B_{0,p}(\theta) = 0,$$

except for $B_{0,0}(\theta) = 1$. Under this form, $B_{k,p}(\theta)$ turns out to be the Bell polynomial in the variables $x_1 = (\theta)_1, x_2 = (\theta)_2, \ldots, x_k = (\theta)_k$. This leads in particular to $B_{k,1}(\theta) = (\theta)_k$, $k \geqslant 1$ and to

$$\mathbf{P}_\theta(P_{n,k} = 1) = \frac{n(\theta)_k}{(n\theta)_k} .$$

Special cases Let us now study the three special cases mentioned in the introduction of this section.

1. **Bose-Einstein case.** When $\theta = 1$, equation (3) simplifies to

$$\mathbf{P}_1(\mathcal{B}_{n,k}(1) = b_1, \ldots, \mathcal{B}_{n,k}(p) = b_p; P_{n,k} = p) = \frac{\binom{n}{p}}{\binom{n+k-1}{k}} ,$$

which is independent of the cell occupancies $(b_1, \ldots, b_p)$ (i.e. the probability is uniform). As there are $\binom{k-1}{p-1}$ sequences $b_q \geqslant 1$ for all $q \in \{1, \ldots, p\}$, satisfying $\sum b_q = k$, we get $B_{k,p}(1) = \frac{k!}{p!}\binom{k-1}{p-1}$ (called Lah numbers) and

$$\forall p \in \{1, \ldots, n \wedge k\} , \quad \mathbf{P}_1(P_{n,k} = p) = \frac{\binom{n}{p}\binom{k-1}{p-1}}{\binom{n+k-1}{k}} .$$

Equation (4) reduces to

$$\mathbf{P}_1(\mathcal{A}_{n,k}(1) = a_1, \ldots, \mathcal{A}_{n,k}(k) = a_k; P_{n,k} = p) = \frac{p!\binom{n}{p}}{\binom{n+k-1}{k}} \frac{1}{\prod_{i=1}^k a_i!} .$$

2. **Maxwell-Boltzmann case.** As $\theta \uparrow \infty$, the probability displayed in equation (3) converges to

$$\mathbf{P}_\infty(\mathcal{B}_{n,k}(1) = b_1, \ldots, \mathcal{B}_{n,k}(p) = b_p; P_{n,k} = p) = \binom{n}{p} \frac{k!}{\prod_{q=1}^p b_q!} \frac{1}{n^k} .$$

With $S_{k,p}$ the second kind Stirling numbers, we get

$$\forall p \in \{1, \ldots, k\} , \quad \mathbf{P}_\infty(P_{n,k} = p) = \frac{n! S_{k,p}}{(n-p)! n^k} . \tag{9}$$

This result is ancient and well-known (see Johnson and Kotz, 1969). Equation (4) convergences to

$$\mathbf{P}_\infty(\mathcal{A}_{n,k}(1) = a_1, \ldots, \mathcal{A}_{n,k}(k) = a_k; P_{n,k} = p) = \frac{n! k! n^{-k}}{(n-p)! \prod_{i=1}^k (i^{a_i} a_i!)} .$$

3. **Kingman case.** Consider the situation where $n \uparrow \infty$, $\theta \downarrow 0$ while $n\theta = \gamma > 0$. In such case, the probability displayed in equation (3) converges to

$$\mathbf{P}^*_\gamma(\mathcal{B}_k(1) = b_1, \ldots, \mathcal{B}_k(p) = b_p; P_k = p) = \frac{k!}{p!} \frac{\gamma^p}{(\gamma)_k \prod_{q=1}^p b_q} .$$

With $s_{k,p}$ the absolute value of the first kind Stirling numbers, we get

$$\forall p \in \{1, \ldots, k\} , \quad \mathbf{P}^*_\gamma(P_k = p) = \frac{\gamma^p s_{k,p}}{(\gamma)_k} . \tag{10}$$

It follows that the probabilities displayed in examples (7) and (8) converge respectively to

$$\frac{\gamma}{\gamma + k} \quad \text{and} \quad \frac{b_r}{\gamma + k} . \tag{11}$$

We note also that the distribution of P_k in this case is in the class of exponential families. We recall the important result of Korwar and Hollander (1973):

$$\frac{P_k}{\log k} \xrightarrow[k \to \infty]{a.s.} \gamma .$$

At least, in the Kingman limit, the probability displayed in (4) converges to

$$\mathbf{P}^*_\gamma(\mathcal{A}_k(1) = a_1, \ldots, \mathcal{A}_k(k) = a_k; P_k = p) = \frac{k! \gamma^p}{(\gamma)_k \prod_{i=1}^k (i^{a_i} a_i!)} .$$

3 Estimation of the number of species

In this section we now investigate several statistical aspects dealing with the estimation of the number of species. We shall start with considering the problem of estimating the number of species, assuming first θ to be known. The proposed procedure to estimate (θ, n) is explained after. Then we consider two stopping rules for the sampling process and a goodness-of-fit procedure. To conclude, numerical simulations were carried out.

3.1 Estimation of n when θ is known

Using theorems 2.1 and 2.2, one can easily derive the conditional distributions of $(\mathcal{B}_{n,k}(1), \ldots, \mathcal{B}_{n,k}(p))$ and of $(\mathcal{A}_{n,k}(1), \ldots, \mathcal{A}_{n,k}(k))$ which are respectively:

$$\mathbf{P}_\theta(\mathcal{B}_{n,k}(1) = b_1, \ldots, \mathcal{B}_{n,k}(p) = b_p \mid P_{n,k} = p) = \frac{k!}{p!}\frac{1}{B_{k,p}(\theta)}\prod_{q=1}^{p}\frac{(\theta)_{b_q}}{b_q!} \tag{12}$$

and:

$$\mathbf{P}_\theta(\mathcal{A}_{n,k}(1) = a_1, \ldots, \mathcal{A}_{n,k}(k) = a_k \mid P_{n,k} = p) = \frac{k!}{B_{k,p}(\theta)}\prod_{i=1}^{k}\frac{(\theta)_i^{a_i}}{i!^{a_i}a_i!} \tag{13}$$

These conditional probabilities being independent of n, it follows that the random variable $P_{n,k}$ is a sufficient statistic.

Assume now that $k > p$. Using log-concavity in n of $\mathbf{P}_\theta(P_{n,k} = p)$, the maximum likelihood estimator $\widehat{n}$ is given implicitly by:

$$\frac{\mathbf{P}_\theta(P_{\widehat{n},k} = P)}{\mathbf{P}_\theta(P_{\widehat{n}-1,k} = P)} = 1.$$

From equation (5), identifying $\widehat{n}$ with the largest integer short of the solution, the estimator $\widehat{n}$ we suggest is the fixed point of:

$$\widehat{n} = P + \widehat{n}\frac{((\widehat{n}-1)\theta)_k}{(\widehat{n}\theta)_k}\ .$$

This estimator is biased from above. The estimator $\widetilde{n}$ of Keener *et al.* (1986) is given by:

$$\widetilde{n} = P + \frac{B_{k,P-1}(\theta)}{B_{k,P}(\theta)}\ . \tag{14}$$

If $k \geqslant n$, it is unbiased attaining the minimum variance bound (UMVB) and in this case we have:

$$\mathbf{E}_\theta\left(\frac{B_{k,P-1}(\theta)}{B_{k,P}(\theta)}\right) = n\frac{((n-1)\theta)_k}{(n\theta)_k} = n\left\langle\theta\right\rangle_{n,k;1}\ .$$

In practice, it is interesting to plot the observed number of species P against sample size k. If $n < \infty$, P should stabilize to an asymptote. If this is not the case, P should drift at ∞ with k. For example, consider the following situation where $P/k \to \rho \in (0,1)$ when $k \to \infty$ and $P \to \infty$. Using an asymptotic representation of $B_{k,p}(\theta)$ in this limit, one gets that $\frac{\widetilde{n}}{k} \to \rho_*$ where $\rho_* > 0$ is defined implicitly (see Keener *et al.* 1986) by:

$$\rho = \rho_*\left(1 - \left(\frac{\theta\rho_*}{1+\theta\rho_*}\right)^\theta\right)\ .$$

Asymptotic normality of $(\widetilde{n} - k\rho_*)/\sqrt{k}$ could be proved as $k \to \infty$.

Special cases

1. **Bose-Einstein case ($\theta = 1$).** We find explicitly:

$$\widehat{n} = \frac{P(k-1)}{k-P} \qquad \text{and} \qquad \widetilde{n} = \frac{Pk}{k-P+1}.$$

The maximal value which $\widehat{n}$ can take is obtained if $k - P = 1$; in this case $\widehat{n} = P^2$. Its minimal value is 1 if $P = 1$ for all k. Note that in the Bose-Einstein model, $\rho_* = \rho/(1-\rho)$ and both $\widehat{n}/k$ and $\widetilde{n}/k$ would converge to ρ_* assuming the asymptotic regime $P/k \to \rho$.

2. **Maxwell-Boltzmann case ($\theta \to \infty$).** The maximum likelihood $\widehat{n}$ solves:

$$P = \widehat{n}\left(1 - \left(1 - \frac{1}{\widehat{n}}\right)^k\right) ,$$

and, with $S_{k,p}$ the second kind Stirling numbers, the UMVB estimator $\widetilde{n}$ in (14) becomes:

$$\widetilde{n} = P + \frac{S_{k,P-1}}{S_{k,P}} ,$$

recalling $B_{k,p}(\theta) \sim \theta^k S_{k,p}$ as $k \to \infty$.

3. **Kingman case.** Indeed there is no estimation of n problem (because $n = \infty$), rather the problem is to estimate $\gamma > 0$ which is the unique remaining parameter. A situation in which the Kingman model fits best to data is a situation for which one should conclude $n = \infty$. Recalling equation (10), the MLE $\widehat{\gamma}$ of γ is characterized by $\partial_\gamma \log \mathbf{P}^*_\gamma(P_k = P)(\widehat{\gamma}) = 0$, hence implicitly by:

$$\xi_k(\widehat{\gamma}) = \sum_{l=0}^{k-1} \frac{\widehat{\gamma}}{\widehat{\gamma} + l} = P . \tag{15}$$

It is biased and involves the problem of inverting the generalized harmonic sequence ξ_k. The properties of this estimator are well studied (see Carlton (1999) for a review). In particular,

$$\widehat{\gamma} \xrightarrow[k\to\infty]{Pr} \gamma .$$

In sharp contrast with the finite n case, there is no UMVB estimator $\widetilde{\gamma}$ of γ itself (nor of any polynomials in γ), because if $\widetilde{\gamma} = \phi(P)$ existed and were unbiased, function ϕ would satisfy:

$$\sum_{p=1}^{k} \gamma^p s_{k,p} \phi(p) = \gamma(\gamma)_k ,$$

which is impossible because the left-hand-side is a polynomial of degree at most k in γ whereas the right-hand-side is a polynomial of degree $k+1$. So, if the problem is to estimate γ, $\widehat{\gamma}$ turns out to be the more satisfactory estimate in this case, despite its biased property. However there are UMVB estimators of rational functions of γ of the form:

$$\forall l \in \{1, \ldots, k\} , \quad r_l(\gamma) = \frac{\gamma(\gamma)_{k-l}}{(\gamma)_k} .$$

They are given by:

$$\widetilde{r_l} = \frac{s_{k-l,P-1}}{s_{k,P}} . \tag{16}$$

For instance, from equation (10), recalling $s_{k,p} = 0$ if $p > k$:

$$\mathbf{E}^*_\gamma\left(\frac{s_{k-l,P-1}}{s_{k,P}}\right) = \frac{\gamma}{(\gamma)_k} \sum_{p=1}^{k-l} \gamma^p s_{k-l,p} = r_l(\gamma) .$$

In particular, when $l = 1$, $\widetilde{r}_1 = \frac{s_{k-1,P-1}}{s_{k,P}}$ is an UMVB estimator of $r_1(\gamma) = \frac{\gamma(\gamma)_{k-1}}{(\gamma)_k} = \frac{\gamma}{\gamma+k-1}$ which, from equation (11), is the probability to observe a new species from k-th trial (see Ewens, 1996).

3.2 Joint estimations of θ and n

In some applications, θ is also unknown and the question of its simultaneous estimation arises. As $P_{n,k}$ is not a sufficient statistic for θ (from equation (12), for example), we turn to a different point of view. We briefly recall the idea of an estimator studied in Huillet and Paroissin (2005). With $(U_1, \ldots, U_k)$ the k iid uniform random sample on $[0,1]$, let $(M_1, \ldots, M_k)$ be the corresponding fragments numbers hit (or animals species). With $l_1, l_2 \in \{1, \ldots, k\}$, let:

$$\delta_{l_1,l_2} = \sum_{m=1}^{n} \mathbf{I}(M_{l_1} = m; M_{l_2} = m) ,$$

denote the event that $M_{l_1} = M_{l_2}$ for some fragment in $\mathbf{S}_n$. Introduce the pair-matching statistic:

$$D_{n,k} = \frac{1}{k(k-1)} \sum_{l_1 \neq l_2 = 1}^{k} \delta_{l_1,l_2} \ .$$

It is the empirical probability that two randomly chosen items of the k-sample are identical. In a genetic context, $D_{n,k}$ is called the homozygosity of the sample (Tavaré, 2004). Note that:

$$D_{n,k} = \frac{1}{k(k-1)} \sum_{q=1}^{P_{n,k}} \mathcal{B}_{n,k}(q)(\mathcal{B}_{n,k}(q)-1) = \frac{k}{k-1} \left(\sum_{q=1}^{P_{n,k}} \left(\frac{\mathcal{B}_{n,k}(q)}{k} \right)^2 - \frac{1}{k} \right) \ .$$

Indeed for each visited species q, we need to count the number $\mathcal{B}_{n,k}(q) - 1$ of returns to q, together with its multiplicity, with $\sum_{q=1}^{P_{n,k}} \mathcal{B}_{n,k}(q) = k$. Note that $D_{n,k}$ is a function of $P_{n,k}$ and of $\mathcal{B}_{n,k}(q)$, $q \in \{1, \ldots, P_{n,k}\}$.

The expectation of δ_{l_1,l_2} is the probability that two fragments chosen at random are the same. From equation (4) with $k = 2$ and $p = 1$, $a_1 = 0$, $a_2 = 1$, we get:

$$\mathbf{E}_\theta(D_{n,k}) = \mathbf{E}_\theta(\delta_{l_1,l_2}) = \frac{1+\theta}{1+n\theta} \ . \tag{17}$$

Assume the observations are $P_{n,k} = P$ and $\mathcal{B}_{n,k}(q) = B_q$ for $q \in \{1, \ldots, P\}$. Then the observed value D of $D_{n,k}$ is:

$$D = \frac{k}{k-1} \left(\sum_{q=1}^{P} \left(\frac{B_q}{k} \right)^2 - \frac{1}{k} \right) \ .$$

Applying the method of moments, θ can be estimated by $(1-D)/(nD-1)$ which is a consistent estimator. Therefore, we propose the following estimators of the pair (θ, n):

$$\widehat{\theta}_1 = \frac{1-D}{\widehat{n}_1 D - 1} \quad \text{and} \quad \widehat{n}_1 = P + \widehat{n}_1 \frac{((\widehat{n}_1 - 1)\widehat{\theta}_1)_k}{(\widehat{n}_1 \widehat{\theta}_1)_k} \ , \tag{18}$$

or:

$$\widetilde{\theta}_1 = \frac{1-D}{\widetilde{n}_1 D - 1} \quad \text{and} \quad \widetilde{n}_1 = P + \frac{B_{k,P-1}(\widetilde{\theta}_1)}{B_{k,P}(\widetilde{\theta}_1)} \ . \tag{19}$$

These estimators are based on the couple of observations (P, D) and depend on which estimator of n itself was chosen. The numerical strategy is to get an implicit equation for $\widehat{n}_1$ (or $\widetilde{n}_1$) by substituting the expression of $\widehat{\theta}_1$ (or $\widetilde{\theta}_1$) as a function of $\widehat{n}_1$ (or $\widetilde{n}_1$) in the second equation, solve it in $\widehat{n}_1$ (or $\widetilde{n}_1$) as a fixed point problem and then deduce the corresponding estimates for θ. Note that the functions involved in this fixed point problem are rational.

An alternative estimation procedure which uses the observable P and cell occupancies $\mathcal{B}_{n,k}(q) = B_q$, $q \in \{1, \ldots, P\}$ is as follows. Consider the Renyi entropy of order $\alpha \geqslant 1$ (Piélou, 1975) defined as follows:

$$\phi_n = \frac{1}{1-\alpha} \log \sum_{m=1}^{n} S_m^\alpha \ .$$

When $\alpha = 2$, $\phi_n = -\log(\sum_{m=1}^n S_m^2)$ is the Simpson index of biodiversity (Simpson, 1949) up to the logarithmic transformation. As α tends to 1, ϕ_n tends to $-\sum_{m=1}^n \log S_m$, the Shannon entropy, which is also an index of biodiversity (Piélou, 1975). Consequently we will rather consider random additive functional of $\mathbf{S}_n$:

$$\phi_n = \sum_{m=1}^{n} h(S_m) \ .$$

Hence with $h(s) = s^2$, it is exactly the Simpson index and with $h(s) = -s \log s$, it is the Shannon index. Below we will only consider the former case. The Simpson index of biodiversity can also be viewed as the size of a size-biased sample fragment from $\mathbf{S}_n$ for which:

$$\mathbf{E}_\theta \left(\sum_{m=1}^{n} S_m^2 \right) = n\mathbf{E}_\theta(S_1^2) = \frac{1+\theta}{1+n\theta} \ .$$

We obtain the same expected value as in equation (17): it was already noticed by Simpson (1949). For a review of various measures of species diversity, see Hubálek (2000). According to lemma 1, $\mathcal{K}_{n,k}(m)/k$ is an estimator of S_m, implying that the quantity:

$$\widehat{\phi}_{n,k} = \sum_{m=1}^{n} h\left(\frac{\mathcal{K}_{n,k}(m)}{k}\right)$$

could be used as an estimator of ϕ_n. Clearly $\widehat{\phi}_{n,k} \xrightarrow[k\to\infty]{d} \phi_n$ and in particular:

$$\mathbf{E}_\theta\left(\widehat{\phi}_{n,k}\right) \xrightarrow[k\to\infty]{} \mathbf{E}_\theta(\phi_n) \ .$$

To be effective this supposes n to be known, which could be not true. If n is unknown, enhancing fragments with small probability to occur, we shall rather consider a quantity based on the sample coverage C (the proportion of seen species in a k-sample):

$$C = \sum_{q=1}^{P_{n,k}} S_q \ .$$

Indeed following Chao and Shen (2003), we can consider:

$$\psi_{n,k} = \sum_{m=1}^{n} \frac{h\left(S'_m\right)}{\pi_m\left(S'_m\right)} \mathbf{I}\left(\mathcal{K}_{n,k}\left(m\right) > 0\right) = \sum_{q=1}^{P_{n,k}} \frac{h\left(S'_q\right)}{\pi_q\left(S'_q\right)}$$

where $S'_q = \frac{S_q}{C}$ for all $q \in \{1, \ldots, P_{n,k}\}$ (in order to have $\sum_{q=1}^{P_{n,k}} S'_q = 1$) and where $\pi_q(S'_q) = 1 - (1 - S'_q)^k$ is the probability to observe fragment q among the $P_{n,k}$ which were effectively observed (see equation (2)). Clearly, we have:

$$\psi_{n,k} \xrightarrow[k\to\infty]{d} \phi_n \ .$$

But $\psi_{n,k}$ involves unknown quantities. Hence, if $P_{n,k}$ fragments are observed, an estimator of $S'_q = S_q/C$ is:

$$\widehat{S}'_q = \frac{\mathcal{B}_{n,k}(q)}{k} \ .$$

It follows that a possible estimator of $\psi_{n,k}$ is:

$$\widehat{\psi}_{n,k} = \sum_{q=1}^{P_{n,k}} \frac{h\left(\frac{\mathcal{B}_{n,k}(q)}{k}\right)}{1 - \left(1 - \frac{\mathcal{B}_{n,k}(q)}{k}\right)^k} \ .$$

Particularizing to $h(s) = s^2$ (Simpson index of diversity),

$$\widehat{\psi}_{n,k} = \sum_{q=1}^{P_{n,k}} \frac{\left(\frac{\mathcal{B}_{n,k}(q)}{k}\right)^2}{1 - \left(1 - \frac{\mathcal{B}_{n,k}(q)}{k}\right)^k} \ ,$$

is such that:

$$\mathbf{E}_\theta(\widehat{\psi}_{n,k}) \xrightarrow[k\to\infty]{} \mathbf{E}_\theta(\phi_n) = \frac{\theta + 1}{n\theta + 1} \ .$$

Assuming the observation is $\widehat{\psi}_{n,k} = \psi$, $(1 - \psi)/(n\psi - 1)$ is also a consistent estimator of θ by application of the asymptotic method of moments. Note that:

$$\psi = \sum_{q=1}^{P} \frac{\left(\frac{B_q}{k}\right)^2}{1 - \left(1 - \frac{B_q}{k}\right)^k} \ ,$$

where $B_q \geqslant 1$ is an observed realization of $\mathcal{B}_{n,k}(q)$ for any $q \in \{1, \ldots, P\}$. hence it involves the observations P and B_q with $q \in \{1, \ldots, P\}$. Therefore an alternative closely related to the two previous estimators (see equations (18) and (19)) for the pair (θ, n) could be:

$$\widehat{\theta}_2 = \frac{1 - \psi}{\widehat{n}_2\psi - 1} \quad \text{and} \quad \widehat{n}_2 = P + \widehat{n}_2 \frac{((\widehat{n}_2 - 1)\widehat{\theta}_2)_k}{(\widehat{n}_2\widehat{\theta}_2)_k} \ ,$$

or:

$$\widetilde{\theta}_2 = \frac{1-\psi}{\widetilde{n}_2\psi - 1} \quad \text{and} \quad \widetilde{n}_2 = P + \frac{B_{k,P-1}(\widetilde{\theta}_2)}{B_{k,P}(\widetilde{\theta}_2)} \ .$$

These estimators are based on the set of observations P and B_q, $q \in \{1, \ldots, P\}$ and depend on which estimator of n itself was chosen.

3.3 Stopping rules

Here we now define three stopping rules for the sampling process. Indeed there is no "objective" stopping rules. Each of these stopping rules are based on some simple and interesting questions that arise naturally in our context. What is the sample size until the first visit to the smallest fragment? How long should one wait until all fragments have been visited (the coupon collector problem)? When the probability to discover a new species is smaller that a given threshold? Clearly the two first questions concern only the situation with $n < \infty$ while the last could be answered for any case.

1. Let $S_{(n)}$ be the smallest fragment among $\mathbf{S}_n$. Let $K_{(n)}$ be the sample size until the first visit to $S_{(n)}$. Then

$$\mathbf{P}\left(K_{(n)} > k \mid \mathbf{S}_n\right) = \left(1 - S_{(n)}\right)^k$$

is the conditional waiting time until the first visit to this fragment. Averaging over the partitions $\mathbf{S}_n$, we obtain

$$\begin{aligned} \mathbf{P}_\theta\left(K_{(n)} > k\right) &= \mathbf{E}_\theta \mathbf{P}\left(K_{(n)} > k \mid \mathbf{S}_n\right) = \mathbf{E}_\theta\left[\left(1 - S_{(n)}\right)^k\right] \\ &= \int_0^1 \mathbf{P}_\theta\left[\left(1 - S_{(n)}\right)^k > s\right] ds = 1 - \int_0^1 \mathbf{P}_\theta\left[S_{(n)} > 1 - s^{\frac{1}{k}}\right] ds. \end{aligned}$$

To evaluate this probability, we thus need to compute the distribution of $S_{(n)}$. We can prove

$$\mathbf{P}_\theta\left(S_{(n)} > s\right) = \frac{\Gamma(n\theta)}{\Gamma(\theta)^n}\phi_{n,\theta}(s)$$

where $\phi_{n,\theta}(s) = h^{*n}_{\theta,s}(t) \mid_{t=1}$ is the n-fold convolution of $t \mapsto h_{\theta,s}(t) = t^{\theta-1}\mathbf{I}(1 \geqslant t > s)$ evaluated at $t = 1$. This distribution could be computed in closed form. In the Bose-Einstein case ($\theta = 1$), the expression simplifies to

$$\mathbf{P}_1(S_{(n)} > s) = (1 - ns)_+^{n-1} \ ,$$

where $x_+ = x \vee 0$ (see Huillet, 2003). As a result, with $k \geqslant 1$,

$$\begin{aligned} \mathbf{P}_1\left(K_n > k\right) &= 1 - \int_0^1 \left(1 - n\left(1 - s^{1/k}\right)\right)_+^{n-1} ds \\ &= 1 - \int_{\left(1-\frac{1}{n}\right)^k}^1 \left(1 - n\left(1 - s^{1/k}\right)\right)^{n-1} ds \\ &= 1 - \frac{k}{n}\left(1 - \frac{1}{n}\right)^{k-1} \int_0^1 x^{n-1}\left(1 + \frac{x}{n-1}\right)^{k-1} dx \\ &= 1 - \frac{k}{n}\left(1 - \frac{1}{n}\right)^{k-1} \sum_{j=0}^{k-1} \binom{k-1}{j} (n-1)^{-j} (n+j)^{-1} \ . \end{aligned}$$

2. Let $K_n^+ = \inf\{k \geqslant n \ ; \ P_{n,k} = n\}$ be the first time that all species are observed in the sample. With $k \geqslant n$, we have

$$\mathbf{P}_\theta(K_n^+ > k) = \mathbf{P}_\theta(P_{n,k} < n) = 1 - \mathbf{P}_\theta(P_{n,k} = n) \ .$$

Recalling equation (6), we obtain

$$\mathbf{P}_\theta(K_n^+ > k) = \frac{1}{(n\theta)_k} \sum_{q=1}^{n-1} (-1)^{q-1} \binom{n}{q} ((n-q)\theta)_k \ .$$

Recalling equation (5), with $k \geqslant n$, this may also be written as follows:

$$\mathbf{P}_\theta(K_n^+ \leqslant k) = \frac{n!}{(n\theta)_k} B_{k,n}(\theta) \ .$$

When $\theta = 1$ and $\theta \uparrow \infty$, this formula further simplifies to give respectively the well-known results:

$$\mathbf{P}_1(K_n^+ \leqslant k) = \frac{\binom{k}{n}}{\binom{n+k-1}{n}} = \frac{\binom{k-1}{n-1}}{\binom{n+k-1}{k}} \ ,$$

and:

$$\mathbf{P}_\infty(K_n^+ \leqslant k) = \frac{n! S_{k,n}}{n^k} \ .$$

3. A last possible stopping rule for the sample is the following: the proceeding with sampling is useless if the estimated probability to obtain a new species $\widehat{r}_1$ (or $\widetilde{r}_1$) is less than some small value ε (say $\varepsilon = 0.01$ for instance). Hence we are interested in the two following sample sizes:

$$\widehat{K}_\varepsilon = \inf\{k \ ; \ \widehat{r}_1 < \varepsilon\} \quad \text{and} \quad \widetilde{K}_\varepsilon = \inf\{k \ ; \ \widetilde{r}_1 < \varepsilon\} \ ,$$

using respectively $\widehat{n}$ or $\widetilde{n}$ for estimating n. When P distinct species have been observed at step k, the probability to get a new species at the $(k+1)$-th trial is

$$r_1 = \frac{(n-P)\theta}{n\theta + k} \ .$$

Using estimators developed previously, we obtain the two following estimates for r_1:

$$\widehat{r}_1 = \frac{(\widehat{n}-P)\theta}{\widehat{n}\theta + k} \quad \text{and} \quad \widetilde{r}_1 = \frac{(\widetilde{n}-P)\theta}{\widetilde{n}\theta + k}$$

(if θ is also unknown, one could replace its estimates). When $\theta = 1$ (Bose-Einstein case), the explicit expressions for $\widehat{n}$ and for $\widetilde{n}$ lead to:

$$\widehat{r}_1 = \frac{\widehat{n}-P}{\widehat{n}+k} = \frac{P(P-1)}{k^2-P} \quad \text{and} \quad \widetilde{r}_1 = \frac{\widehat{n}-P}{\widehat{n}+k} = \frac{P(P-1)}{k^2+1} \ .$$

Obviously $\widehat{r}_1 \geqslant \widetilde{r}_1$. As a consequence, $\widehat{K}_\varepsilon \geqslant \widetilde{K}_\varepsilon$. Thus, $\widehat{K}_\varepsilon$ is of order $P\varepsilon^{-1/2}$ and $\widetilde{K}_\varepsilon$ is of order $(P(P-1))^{1/2}\varepsilon^{-1/2}$. Hence if P is large enough, $\widehat{K}_\varepsilon$ and $\widetilde{K}_\varepsilon$ are of the same order. In the case of Kingman model, we have only an explicit expression for $\widetilde{n}$. In such case, let us recall that:

$$\widetilde{r}_1 = \frac{s_{k-1,P-1}}{s_{k,P}} \ ,$$

which could be evaluated from inspection of a table of the first kind Stirling numbers.

3.4 Goodness of fit using the second Ewens sampling formula

Deciding which model fits the best to a concrete situation is a challenging problem. This can first be appreciated from the likelihood of the observations under the different models to be compared. We shall recall an additional procedure followed by Keener *et al.* (1987) for the case $n < \infty$: First, a simple computation of $\alpha_i = \mathbf{E}_\theta(\mathcal{A}_{n,k}(i))$ gives, using our notations

$$\alpha_i = ni \binom{k}{i} \langle\theta\rangle_{n,k-i;1} \ .$$

According to theorem 2.5 in Keener *et al.* (1987), a UMVB estimator of α_i is obtained under the form:

$$\widetilde{\alpha}_i = \mathbf{E}_\theta(\mathcal{A}_{n,k}(i) \mid P) = \frac{(\theta)_i}{(i-1)!} \frac{B_{k-i,P-1}(\theta)}{B_{k,P}(\theta)} \ .$$

When $\theta = 1$ (Bose-Einstein case), it becomes:

$$\widetilde{\alpha}_i = \frac{ip(p-1)}{k-i+1} \frac{(k-i-1)!(k-p)!}{k!(k-1)!} \ ,$$

recalling the expression of $B_{k,P}(1)$. Define next the MLE $\widehat{\alpha}_i$ of α_i to be:

$$\widehat{\alpha}_i = \widehat{n} i \binom{k}{i} \langle \theta \rangle_{\widehat{n},k-i;1} .$$

Based on the observations A_i of $\mathcal{A}_{n,k}(i)$, the goodness of fit of the model can be measured by one of the two following statistics:

$$\widetilde{\chi}^2 = \sum_{i=1}^{k} (A_i - \widetilde{\alpha}_i)^2 / \widetilde{\alpha}_i \qquad \text{or} \qquad \widehat{\chi}^2 = \sum_{i=1}^{k} (A_i - \widehat{\alpha}_i)^2 / \widehat{\alpha}_i .$$

In the case of the Kingman model, one can check that:

$$\alpha_i = \mathbf{E}^*_\gamma(\mathcal{A}_k(i)) = \frac{k!}{i(k-i)!} \frac{\gamma}{\gamma + i - 1}$$

and the second statistic becomes:

$$\widehat{\chi}^2 = \sum_{i=1}^{k} (A_i - \widehat{\alpha}_i)^2 / \widehat{\alpha}_i ,$$

where $\widehat{\alpha}_i = \frac{k!}{(k-i)!} \frac{\widehat{\gamma}}{i(\widehat{\gamma}+i-1)}$ and where $\widehat{\gamma}$ is given by equation (15).

3.5 Numerical simulations

We now apply the estimators developed and discussed previously on simulated data to observe the behavior of their quality when n, θ and k are varying.

We consider empirical distribution of k-samples for a given Dirichlet partition (however one could rather prefer to consider empirical distribution of Dirichlet partitions and one k-sample). For $n \in \{100; 200; 500\}$ and $\theta \in \{\frac{1}{2}; 1; \frac{3}{2}\}$, we simulated a Dirichlet partition. Over this partition, we simulated 500 k-samples with $k \in \{\frac{2n}{3}; n; \frac{3n}{2}\}$. Note that we managed to use the same uniform random variables, so that $P_{n,\frac{3n}{2}}$ corresponds to the same observations than the first ones of $P_{n,n}$ and so on. We considered the two different cases: θ known and θ unknown. In some cases, we did not use the estimators based on $\widetilde{n}$ since computations were too heavy.

Tables 1 to 3 contain the estimations (first when θ is known and then when θ is unknown) respectively for $\theta = 1$, $\theta = \frac{1}{2}$ and $\theta = \frac{3}{2}$. The numbers that appear in the cells are the empirical averages of the estimations of n or θ, and the number in brackets within the cells the empirical standard deviations (over 100 k-samples as described above). Note that comparing standard deviations for the estimations of n and θ does not make sense (one should rather use for instance the coefficient of variation which is dimension-less).

For $\theta = 1$ (table 1), the results are very good, even when θ is unknown (but except for the estimation of θ with the statistic $\Psi_{n,k}$). For $\theta = \frac{1}{2}$ and $\theta = \frac{3}{2}$, we did not run the estimators based on $\widetilde{n}$ for the reason given above. Results are quite good when θ is known, but not so good when θ is unknown.

[Tab. 1 about here.]

[Tab. 2 about here.]

[Tab. 3 about here.]

4 Applications to real data

We applied our estimators to fourteen different sets of real data. These data are of various nature as we will see later. However we will only consider here two kind of real data sets. The first one was studied by Keener *eta al.* (1987) and deals with word usage by two different authors. The interest of this first data set is that we indeed known the number n to be estimated. The second one was extracted from observations made by Janzen (1973): these data correspond to beetles species observed either during the day or during the night and at different season.

4.1 Federalist papers data

These data were considered by Mosteller and Wallace (1984) and concern word usage by James Madison and Alexander Hamilton. *The Federalist* papers were written between 1787 and 1788 to promote the new Constitution of the State of New-York. Published in various newspapers, these papers was signed under a pseudonym (as for instance *Publius*). Each paper was written by one of the three following persons: James Madison, Alexander

Hamilton and John Jay. The author of most of the seventy-seven papers is clearly identified but Madison and Hamilton disputed the authorship of twelve of them. Hence in order to determine the author of these disputed papers, many researchers studied papers written surely by Madison and Hamilton. In particular some of them focused on the occurrences of function words as defined by the Miller-Newman-Friedman list. Mosteller and Wallace (1984) developed a Bayesian approach to solve this problem. In order to do so, they divided a set of well identified texts (either by Madison or by Hamilton) of equal length. It corresponds to the data presented below.

For $q \in \{0, \ldots, 6\}$, the two tables below gives the number $A(q)$ of manuscripts (of the same type as those published in *The Federalist* and with comparable length) in which a specific word ('may' for Madison and 'can' for Hamilton) occurs exactly q times (these two words were the one selected by Keener *et al.* (1987), but Mosteller and Wallace (1984) studied more words). In this case, the exact numbers of manuscripts are known (respectively $n = 262$ and $n = 247$) and so we will be able to compare it with our estimation. Since we have this additional information, tables 4 and 5 contain the column for $q = 0$, which is not available in real applied context.

For these two data sets, Keener *et al.* (1987) computed the estimation of n for the three special cases considered here (i.e. for the Bose-Einstein, Maxwell-Boltzmann and the Kingman models) and also for the case where both n and θ are unknown. Results are compared throughout the log-likelihood. Indeed, in their paper, there is no theoretical development when n and θ are unknown.

- *Madison data:* the sample size is $k = 172$ and the number of distinct kinds of manuscripts is $p = 106$. When using the statistic $D_{n,k}$, we obtain $\widehat{n}_1 = 274.6$ and $\widehat{\theta}_1 = 1.09$. When using the statistic $\widehat{\psi}_{n,k}$, we obtain $\widehat{n}_2 = 274.6$ and $\widehat{\theta}_2 = 0.32$. When assuming that both n and θ are unknown, Keener *et al.* (1987) obtained respectively 217 and 1.998 as estimated values. This value for n is far from its correct value.

[Tab. 4 about here.]

- *Hamilton data:* the sample size is $k = 139$ and the number of distinct kinds of manuscripts is $p = 90$. When using the statistic $D_{n,k}$, we obtain $\widehat{n}_1 = 253.5$ and $\widehat{\theta}_1 = 0.85$. When using the statistic $\widehat{\psi}_{n,k}$, we obtain $\widehat{n}_2 = 4526.3$ and $\widehat{\theta}_2 = 0.01$. The second value is unsatisfactory. However when assuming that both n and θ are unknown, Keener *et al.* (1987) obtained respectively 10,000,001 and 1.094×10^{-5} as estimated values! This value for n is strongly far from its correct value.

[Tab. 5 about here.]

4.2 Tropical insect data

Janzen (1973) observed tropical insects in twenty-five different sites in Costa Rica and the Caribbean Islands. This paper contains a remarkable collection of data. From it, we extracted three series corresponding to beetles collected either in day-time or in night-time, all during a dry season. These data were collected at the same site referred as "Osa secondary" in Janzen (1973). Observations of the first series were collected during the dry season of the year 1967 in day-time while the ones of the second series were collected at the same period in night-time. At least observations of the last series were collected during the dry season of the year 1968 in day-time.

- *Osa secondary/day/dry/1967:* it was observed $k = 996$ beetles and $p = 140$ distinct species. When using the statistic $D_{n,k}$, we obtain $\widehat{n}_1 = 162.7$ and $\widehat{\theta}_1 = 0.219$. When using the statistic $\widehat{\psi}_{n,k}$, we obtain $\widehat{n}_2 = 162.7$ and $\widehat{\theta}_2 = 0.211$.

[Tab. 6 about here.]

- *Osa secondary/night/dry/1967:* it was observed $k = 835$ beetles and $p = 151$ distinct species. When using the statistic $D_{n,k}$, we obtain $\widehat{n}_1 = 184.1$ and $\widehat{\theta}_1 = 0.268$. When using the statistic $\widehat{\psi}_{n,k}$, we obtain $\widehat{n}_2 = 184.1$ and $\widehat{\theta}_2 = 0.252$.

[Tab. 7 about here.]

- *Osa secondary/day/dry/1968:* it was observed $k = 807$ beetles and $p = 143$ distinct species. When using the statistic $D_{n,k}$, we obtain $\widehat{n}_1 = 173.6$ and $\widehat{\theta}_1 = 0.111$. When using the statistic $\widehat{\psi}_{n,k}$, we obtain $\widehat{n}_2 = 173.6$ and $\widehat{\theta}_2 = 0.108$.

[Tab. 8 about here.]

For all the three data sets, the two values of $\widehat{n}$ are identical. Moreover the two values of $\widehat{\theta}$ are close, which is not always the case. It may be due to the fact that many species are abundant.

4.3 Conclusion

These two families of data sets give some illustration of the results obtained when applying the estimators developed in this paper . In fact it also shows the computational limit of them. In particular one can observe that values of the two estimators $\widehat{n}_1$ and $\widehat{n}_2$ of n (respectively based on $D_{n,k}$ and $\widehat{\psi}_{n,k}$) may differ. This should arise especially when most of species are rare and when there were only few abundant species. However, over the fourteen sets of real data we used, this situation occurs four times. Estimations for the three data about tropical beetles seem to be exceptionally satisfactory. It may be due to the presence of many abundant species.

References

Carlton, M. (1999). *Applications of the two-parameter Poisson-Dirichlet distribution*. Ph.D. thesis, University of California, Los Angeles.

Chao, A. and T.-S., S. (2003). Nonparametric estimation of shannon's index of diversity when there are unseen species in sample. *Environ. Ecol. Stat.*, **10**, 429–443.

Donnelly, P. (1986). Partition structures, pòlya urns, the ewens sampling formula and the age of alleles. *Theor. Popul. Biol.*, **30**, 271–288.

Ewens, W. (1972). The sampling theory of selectively neutral alleles. *Theor. Popul. Biol.*, **3**, 87–112.

Ewens, W. (1990). Population genetics theory - the past and the future. In S. Lessard, editor, *Mathematical and Statistical Developments of Evolutionary Theory*. Kluwer, Dordrecht.

Ewens, W. (1996). Some remarks on the law of succession. In N. Y. Springer, editor, *Athens Conference on Applied Probability and Time Series Analysis (1995), Vol. 1*, volume 114 of *Lectures Notes in Statistic*, pages 229–244.

Feller, W. (1966). *Introduction to probability theory and its applications*. Wiley, New-York.

Hubálek, Z. (2000). Measures of species diversity in ecology: an evaluation. *Folia Zool.*, **49**(4), 241–260.

Huillet, T. (2003). Sampling problems for randomly broken sticks. *J. Phys. A*, **36**(14), 3947–3960.

Huillet, T. (2005). Sampling formulae arising from random dirichlet populations. *Communications in Statistics - Theory and Methods*, **34**(5), 1019–1040.

Huillet, T. and Paroissin, C. (2005). Estimation of the parameter of a dirichlet distribution using residual allocation representations and sampling properties. *Stat. Meth.*, **2**(2), 95–110.

Janzen, D. H. (1973). Sweep samples of tropical foliage insects: description of study sites, with data on species abundances and size distributions. *Ecology*, **54**, 659–686.

Johnson, N. L. and Kotz, S. (1969). *Distributions in statistics: Discrete distributions*. Houghton Mifflin Co., Boston, Mass.

Keener, R., Rothman, E., and Starr, N. (1987). Distributions on partitions. *Ann. Stat.*, **15**(4), 1466–1481.

Kingman, J. (1975). Random discrete distributions. *J. R. Stat. Soc., Ser. B*, **37**, 1–22.

Korwar, R. M. and Hollander, M. (1973). Contributions to the theory of dirichlet processes. *Ann. Probab.*, **1**, 705–711.

McArthur, R. (1957). On the relative abundance of bird species. *Proc. Nat. Acad. Sci. U.S.A.*, **43**, 293–295.

Mosteller, F. and Wallace, D. (1984). *Applied Bayesian and classical inference: the case of the Federalist papers*. Springer-Verlag, New-York.

Piélou, E. (1975). *Ecological diversity*. John Wiley, New-York.

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

Tavaré, S. (2004). Ancestral inference in population genetics. In J. Picard, editor, *Lectures on Probability Theory and Statistics (École d'été de Saint-Flour XXXI, 2001)*, volume 1837 of *Lectures Notes in Mathematics*. Springer, Berlin.

Tokeshi, M. (1993). Species abundance patterns and community structure. *Adv. Ecol. Res.*, **24**, 112–186.

Vlad, M. O., Tsuchiya, M., Oefner, P., and Ross, J. (2001). Bayesian analysis of systems with random chemical composition: renormalization-group approach to dirichlet distributions and the statistical theory of dilution. *Phys. Rev. E*, **65**, 1–8.

Wilks, S. (1962). *Mathematical statistics*. Wiley, New-York.

		θ known		θ unknown			
		$\widehat{n}$	$\widetilde{n}$	$(\widehat{n}_1, \widehat{\theta}_1)$ based on $D_{n,k}$		$(\widehat{n}_2, \widehat{\theta}_2)$ based on $\psi_{n,k}$	
$n = 100$	$k = 66$	**92.96**	90.98	1.2E+09	0.21	9.8E+08	2.7E-08
		(16.51)	(15.80)	(1.3E+09)	(3.5E+08)	(0.32)	(1.1E-08)
	$k = 100$	**93.69**	**92.81**	2821.86	0.21	3322.34	0.01
		(13.29)	(13.05)	(1884.31)	(12632.13)	(0.33)	(0.01)
	$k = 150$	**91.83**	91.46	**91.83**	**1.13**	130.06	**0.61**
		(9.62)	(9.55)	(9.63)	(268.89)	(0.23)	(0.12)
$n = 200$	$k = 133$	206.76	**204.33**	**223.08**	**1.09**	4731.03	0.01
		(31.23)	(30.46)	(153.28)	(0.23)	(767.15)	(0.001)
	$k = 200$	**201.22**	**200.21**	**201.22**	**1.11**	**201.22**	0.46
		(18.07)	(17.89)	(18.07)	(0.18)	(18.07)	(0.04)
	$k = 300$	**200.30**	**199.85**	**200.30**	**1.09**	**200.30**	0.58
		(13.28)	(13.22)	(13.28)	(0.15)	(13.28)	(0.05)
$n = 500$	$k = 333$	**513.34**	**510.96**	**513.34**	**0.99**	**513.34**	0.32
		(41.89)	(41.49)	(41.89)	(0.13)	(41.89)	(0.02)
	$k = 250$	**515.20**	**514.14**	**515.20**	**0.99**	**515.20**	0.43
		(27.79)	(27.68)	(27.79)	(0.13)	(27.79)	(0.03)
	$k = 750$	**515.15**	**514.68**	**515.15**	**0.98**	**515.15**	0.54
		(21.61)	(21.57)	(21.61)	(0.10)	(21.61)	(0.03)

Tab. 1: Estimation over simulated data with $\theta = 1$

		θ known	θ unknown			
		$\widehat{n}$	$(\widehat{n}_1, \widehat{\theta}_1)$ based on $D_{n,k}$		$(\widehat{n}_2, \widehat{\theta}_2)$ based on $\psi_{n,k}$	
$n = 100$	$k = 66$	**108.21**	8.2E+08	0.16	7.5E+08	3.2E-08
		(19.39)	(8.5E+08)	(0.30)	(2.7E+08)	(1.3E-08)
	$k = 100$	**110.36**	1418.81	**0.29**	1708.36	0.02
		(13.93)	(1547.26)	(0.32)	(968.16)	(0.02)
	$k = 150$	162.34	**156.51**	0.91	2462.54	**0.38**
		(15.58)	(614.64)	(0.23)	(16084.16)	(0.29)
$n = 200$	$k = 133$	136.75	4.2E+08	0.33	5.9E+08	0.002
		(84.83)	(6.9E+08)	(0.41)	(3.8E+08)	(0.005)
	$k = 200$	141.06	836.19	**0.42**	1222.67	0.11
		(87.25)	(1328.58)	(0.38)	(1371.98)	(0.19)
	$k = 300$	**183.38**	**473.93**	**0.69**	2828.47	0.19
		(69.20)	(3672.28)	(0.310)	(30533.87)	(0.26)
$n = 500$	$k = 333$	232.11	4.2E+08	0.35	5.9E+08	0.07
		(270.59)	(6.9E+08)	(0.39)	(3.8E+08)	(0.15)
	$k = 250$	234.31	882.81	**0.41**	1269.30	0.11
		(272.37)	(1308.03)	(0.37)	(1338.81)	(0.18)
	$k = 750$	**277.95**	**521.22**	**0.68**	2875.75	0.19
		(256.56)	(3669.47)	(0.30)	(30529.92)	(0.25)

Tab. 2: Estimation over simulated data with $\theta = \frac{1}{2}$

		θ known	θ unknown			
		$\widehat{n}$	$(\widehat{n}_1, \widehat{\theta}_1)$ based on $D_{n,k}$		$(\widehat{n}_2, \widehat{\theta}_2)$ based on $\psi_{n,k}$	
$n = 100$	$k = 66$	**99.39**	7.801E+08	0.18	8.299E+08	0.0001
		(19.49)	(1.167E+09)	(0.28)	(3.365E+08)	(0.003)
	$k = 100$	**99.37**	1740.15	0.26	1674.75	0.01
		(13.48)	(2273.91)	(0.30)	(1331.18)	(0.01)
	$k = 150$	155.01	**465.64**	**0.74**	2820.17	0.0001
		(18.25)	(3673.02)	(0.37)	(30534.60)	(0.25)
$n = 200$	$k = 133$	151.57	4.195E+08	0.40	5.922E+08	0.002
		(110.93)	(6.926E+08)	(0.47)	(3.830E+08)	(0.005)
	$k = 200$	151.08	853.30	0.47	1239.78	0.09
		(107.15)	(1320.33)	(0.45)	(1359.18)	(0.15)
	$k = 300$	**189.47**	**488.62**	**0.75**	2843.15	0.19
		(81.03)	(3671.15)	(0.38)	(30532.61)	(0.25)
$n = 500$	$k = 333$	254.38	4.2E+08	0.37	5.9E+08	0.05
		(314.97)	(6.9E+08)	(0.41)	(3.8E+08)	(0.11)
	$k = 250$	253.11	921.32	0.44	1307.81	0.09
		(309.82)	(1295.86)	(0.40)	(1315.75)	(0.15)
	$k = 750$	**291.79**	**556.83**	**0.71**	2911.36	0.18
		(284.10)	(3668.96)	(0.33)	(30527.13)	(0.24)

Tab. 3: Estimation over simulated data with $\theta = \frac{3}{2}$

q	0	1	2	3	4	5	6
$A(q)$	156	63	29	8	4	1	1

Tab. 4: Madison data

q	0	1	2	3	4	5	6
$A(q)$	157	60	20	5	2	2	1

Tab. 5: Hamilton data

q	1	2	3	4	5	6	7	8	9	10	11	12	14
$A(q)$	70	17	4	5	5	5	5	3	1	2	3	2	2

q	17	29	20	21	24	26	40	57	60	64	71	77	
$A(q)$	1	2	3	1	1	1	1	2	1	1	1	1	

Tab. 6: Osa secondary/day/dry/1967 data

q	1	2	3	4	5	7	8	9	10	11	12
$A(q)$	61	24	13	12	5	6	5	2	4	2	3
q	13	15	17	18	19	26	30	33	40	44	62
$A(q)$	1	1	1	2	2	1	1	1	1	1	2

Tab. 7: Osa secondary/night/dry/1967 data

q	1	2	3	4	5	6	7	9	10	11	12	13
$A(q)$	85	12	10	4	6	3	5	1	2	1	1	1
q	15	18	20	24	25	28	29	30	79	106	112	
$A(q)$	1	2	1	1	1	1	1	1	1	1	1	

Tab. 8: Osa secondary/day/dry/1968