A new estimator for the number of species in a population

Lorenzo Cecconi and Alberto Gandolfi

Università di Firenze, Firenze, Italy

Chelluri C. A. Sastri

Missouri University of Science and Technology, Rolla, MO, USA
and Dalhousie University, Halifax, Canada

Abstract

We consider the classic problem of estimating $T$, the total number of species in a population, from repeated counts in a simple random sample. We first show that the frequently used Chao-Lee estimator can in fact be obtained by Bayesian methods with a Dirichlet prior, and then use such clarification to develop a new estimator; numerical tests and some real experiments show that the new estimator is more flexible than existing ones, in the sense that it adapts to changes in the normalized interspecies variance $\gamma^2$. Our method involves simultaneous estimation of $T$, $\gamma^2$, and of the parameter $\lambda$ in the Dirichlet prior, and the only limitation seems to come from the required convergence of the prior which imposes the restriction $\gamma^2 \leq 1$. We also obtain confidence intervals for $T$ and an estimation of the species’ distribution. Some numerical examples are given, together with applications to sampling from a Census database closely following Benford’s law, showing good performances of the new estimator, even beyond $\gamma^2 = 1$. Tests on confidence intervals show that the coverage frequency appears to be in good agreement with the desired confidence level.

AMS (2000) subject classification. Primary 62G05; Secondary 62P10, 62P30, 62P35.

Keywords and phrases. Bayesian posterior, confidence interval, Dirichlet prior, point estimator, simple random sample, unobserved species, unobserved probability.

1 Introduction

We consider the classic problem of estimating the number $T$ of species in a population, and, subsequently, their distribution, from a simple random sample drawn with replacement. We are interested in the “small sample” regime in which it is likely that not all species have been observed, a question
A new estimator for the number of species

arising in very many situations: see, for instance, Böhning et al. (2004), Brose, Martinez and Williams (2003), Bunge and Fitzpatrick (1993), Church and Gale (1991), Efron and Thisted (1976), Esty (1986), Gandolfi and Sastri (2004), Shen, Chao and Lin (2003); see also http://viceroy.eeb.uconn.edu/estimates. In particular, Bunge and Fitzpatrick (1993) calls for some development of a Bayesian method for the estimation of \( T \), which is the direction that we end up taking.

Related to the estimation of \( T \) is the problem of estimating the total probability \( U \) of the unobserved species, which is also the probability of observing a new species on the next trial, and many, often unrelated, methods have appeared in the literature to deal with both problems. Our first result is that two of the most well-known methods for estimating \( U \), the Turing-Good (Good 1953) and a suitable modification of Laplace’s add-\( \lambda \) (Orlitsky, Santhanam and Zhang, 2003) estimators, coincide in their prediction if and only if \( T \) equals the most well known among its own estimators, namely that of Chao-Lee (Chao and Lee, 1992). Such a connection gives a new meaning to the normalized interspecies variance \( \gamma^2 \), which is a quantity whose evaluation is taken as a preliminary step in the Chao-Lee method, in terms of \( \lambda \). Actually, it suggests that we consider a simultaneous estimation of \( T \) and either \( \gamma^2 \) or \( \lambda \), which is what we sketch in Section 3. Even if this shows some improvement upon previously known methods, it is quickly clear that there is still something missing (as predicted in Esty, 1986) and that a more thorough understanding of the problem is needed.

Although it is not often realized, Laplace’s add-\( \lambda \) is actually based on the Bayesian posterior distribution on the species probabilities starting from a Dirichlet prior (see Good, 1965; Jeffreys, 1961; Johnson, 1932; Zabell, 1982 for the use of such prior). Because of the previously discovered connections, the entire issue can now be phrased in particular Bayesian terms; see also Boender and Rinnoy Kan (1987), Marchand and Schroeck (1982) for unrelated Bayesian approaches to this problem. This is the main content of our work, and it is carried out in Section 4. Eventually, we determine a new estimator \( \hat{T}_{\lambda} \), which performs rather well in several examples.

The main mathematical difficulties in developing the new method concern the admissibility of the parameters. In the first place, \( \gamma^2 \) is nonnegative, but the corresponding solution to the equations leading to its simultaneous estimation with \( T \) might not be so. Therefore, we replace, already in Section 3, one of the equations by a minimization. In the second place, we need to guarantee that for any sample \( \hat{T}_{\lambda} \) is not less than the number \( N \) of observed species, and that the estimate \( \hat{\lambda} \) of \( \lambda \) is not less than 1, as required.
for a proper Dirichlet prior to exist. We are able to prove that all these consistency results hold, as shown in Section 4.

Notice that for many samples the solution $\lambda$ of the minimization problem is one of the extreme values $1$ and $\infty$, so that these two values occur quite often, a phenomenon observed but not yet explained in the literature (see Esty, 1985 for a discussion of this point).

Next, notice that the requirement that $\lambda \geq 1$, which in our scheme is equivalent to $\gamma^2 \leq 1$, is the only real limitation of the method and that there are skewed distributions with $\gamma^2 > 1$ (see Section 6). Yet, it appears that our estimation method works well even for a range of values of $\gamma^2$ exceeding $1$, even up to $3$; it is only for very large values of $\gamma^2$ that the our estimators fails, but this limitation seems to be shared by all existing estimators. In fact, the ACE (Abundance-based Coverage Estimator) method was developed to deal with $\gamma^2 > 1$ (Chao, 2004; Chao, Ma and Yang, 1993; Chao et al., 2000): the method separates “abundant” and “rare” species as those appearing more or less than a fixed $K$, applies usual estimation to rare species (presumably with a reduced $\gamma^2$) and then adds the number of the abundant ones. Upon the suggestion of a referee, we have made extensive tests about using ACE, applied in particular to the Chao-Lee estimator and to ours. It appears that this method is effective up to $\gamma^2 \approx 3$, in which our method already performs well, so we got only a moderate improvement. For much larger values of $\gamma^2$ nothing seems to work.

In addition, in Section 5, we use the estimates of $T$ and $\lambda$ to provide estimates of the species probabilities, for both the observed and the unobserved species. This allows us also to generate confidence intervals for $T$ by means of resampling from the estimated distribution.

Finally, in Section 6 we present some numerical simulations which suggest that $\hat{T}_\lambda$ is more flexible than existing estimators. The performance of all estimators seems to greatly depend on the normalized variance $\gamma^2$, and for different values of $\gamma^2$ different estimators achieve the best estimation suggesting that each has its own merits. Nonetheless, the new estimator seems flexible in the sense that it appears to have acceptable performances for all values of $\gamma^2 \in [0, 3]$. We also present a real test on discovering the number of digits from a data set, the census data in the example. This type of data is known to closely follow Benford’s law (see Benford, 1938; Hill, 1995; and Tao, 2009) for groups of digits of any size. We simulate the sequential discovery of the number of $T = 90$ two-digit pairs in our numerical system from a census data base, and compare our estimator with that of Chao-Lee and their ACE versions. Next, the estimates of the species probabilities obtained in
Section 5 are used to improve the empirical estimates of Benford’s distribution, showing excellent improvement. In the end, we test the confidence intervals for \( T \) and find very good agreement between the confidence level and the fraction of samples in which the interval contains the true value.

2 Some known estimators of \( T \) and related quantities

To fix the notation, assume that the population has a total of \( T \) species having proportions \( p_1, \ldots, p_T \), and that in a sample \( x_1, \ldots, x_n \) of size \( n \) there are \( N \) observed species. For \( i = 1, \ldots, T \), let \( m_i \) be the number of observations of the species \( i \) in the sample, so that \( \sum_{i=1}^{N} m_i = n \). We assume that the \( m_i \)'s are given one of the possible orders in which \( m_1 \geq m_2 \ldots, m_N \geq 1 \) and \( m_i = 0 \) for \( i = N + 1, \ldots, T \). Also, for \( j = 1, \ldots, n \), let \( n_j \) be the prevalence of \( j \), which is to say the number of species observed exactly \( j \) times, so that \( \sum_{j=1}^{n} n_j = N \). Next, let \( L_n(i) = m_i/n \) be the empirical frequency of species \( i \), so that \( C = \sum_{i:L_n(i)>0} p_i \) is the coverage, i.e, the total probability of the observed species, and \( U = 1 - C = \sum_{i:L_n(i)=0} p_i \) is the unobserved probability. We are interested in the estimation of \( T \) and the \( p_i \)'s.

If only interested in estimating the probability \( U \) of observing a new species at the next trial, one may assume that there are \( N + 1 \) species, the \( N \) observed ones and the “new” species with probability \( U \) (see, e.g., Lijoi, Mena and Prünster, 2007; Mao, 2004), which suggests a close relationship between the estimations of \( U \) and \( T \) (as in the title of Esty, 1985). Under this assumption, interesting estimators of the \( p_i \)'s are Laplace’s “add-\( \lambda \)” (Laplace, 1995; Orlitsky et al., 2003):

\[
\hat{p}_i = \frac{m_i + \lambda}{\lambda + \sum_{i \geq 1} (m_i + \lambda)} = \frac{m_i + \lambda}{n + (N + 1)\lambda},
\]

thus estimating \( U \) by \( \hat{U}_{L,\lambda} = \lambda/n + (N + 1)\lambda \), and the abundance-based estimator of Turing and Good (1953), \( \hat{p}_i = 1/n_j(j + 1)n_{j+1}/n \), if \( i \) appeared \( j \geq 0 \) times; since we are taking \( n_0 = 1 \), this gives

\[
\hat{U}_{TG} = \frac{n_1}{n}.
\]

According to Orlitsky et al. (2003), which evaluates estimators of \( U \) by introducing attenuation, \( \hat{U}_{TG} \) is quite effective since it is simple but still with low attenuation; the asymptotic performance could be improved at the cost of more involved expressions, so to keep things simple we adopt \( \hat{U}_{TG} \) in the rest of the paper. We also disregard the smoothing proposed
in Good (1953), aimed at removing isolated zeros in the $n_j$’s, which could certainly give slight improvements but would add annoying details to the presentation. In particular, it would change $n_1 = 0$ into $n_1 = 1$, and we use here the former condition as a simple indicator that all species have been observed.

Returning to estimators of $T$, our work relates to the nonparametric methods in Church and Gale (1991), Harris (1968), and in particular to the frequently used Chao-Lee estimator (Chao and Lee, 1992)

$$\hat{T}_{CL}(\hat{\gamma}) = \frac{N}{1 - \hat{U}_{TG}} + \frac{n\hat{U}_{TG}}{(1 - \hat{U}_{TG})}\hat{\gamma}^2,$$

with $\hat{\gamma}^2$ an estimate of the normalized variation coefficient $\gamma^2$ of the $p_i$’s. A more precise meaning can be given to $\gamma^2$ by assuming that $p$ is itself a random variable uniformly distributed on the $T$ population probabilities $p_1, \ldots, p_T$; in this case, its average is $\bar{p} = T^{-1}\sum_{k=1}^{T} p_k = T^{-1}$, and its normalized variation coefficient becomes

$$\gamma^2 = \frac{\text{var}(p)}{[E(p)]^2} = T\sum_{k=1}^{T}(p_k - \bar{p})^2 = T\sum_{k=1}^{T} p_k^2 - 1.$$  \hspace{1cm} (2.3)

Note that the expression (2.2) appeared in Esty (1985) with $\gamma^2 = 1/k$, where a negative binomial prior with parameter $k$ was considered and then an estimation of $k$ was needed; furthermore it appeared in Böhlne and Schön (2005), Bishop, Fienberg and Holland (1975), Darroch and Ratcliff (1980), Lindsay and Roeder (1987), with $\gamma = 0$ to deal with the uniform case; in the sequel we denote $\hat{T}_{Unif} = \hat{T}_{CL}(0)$.

To get an estimate of $\gamma^2$, Chao and Lee (1992) proceed from (2.3) by using an estimate of Good and Toulmin (1956)

$$\sum_{k=1}^{T} p_k^2 \approx \hat{V}_{GT} = \sum_{j \geq 1} j(j - 1)n_j \frac{n(n - 1)}{n(n - 1)}$$  \hspace{1cm} (2.4)

and using one preliminary estimate for $T$, $\hat{T}_{Unif}$ for instance, to obtain an expression for $\hat{\gamma}^2$. It appears, however, that for some populations $\hat{T}_{CL}$ is converging quite slowly to the true value (see Section 6, in particular figure 3) and we believe that this is due to the errors in the preliminary evaluation of $T$.

To correct the sensitivity to initial values, one could attempt a simultaneous estimation of $T$ and $\gamma^2$; this is mentioned in Esty (1986), hinting at a
second moment method, but deemed too unstable (in particular, Esty, 1986, Section 4, reports on experiments showing an extremely high variance). We now briefly indicate a way of performing the simultaneous estimation which yields quite good results except for some particular samples. Equations (2.2) and (2.3), with \( \sum_{k=1}^{T} p_k^2 \) replaced by \( \hat{V}_{GT} \) as given in (2.4), can be considered simultaneously in the unknowns \( T \) and \( \gamma^2 \), searching for solutions \( \hat{T} \geq N \) and \( \hat{\gamma}^2 \geq 0 \). However, simple examples show that sometimes \( \hat{\gamma}^2 < 0 \): this happens, for instance, when all species are observed once except for one observed three times, as one can easily verify. We then suggest replacing equation (2.3) by a minimization:

\[
\hat{\gamma}^2 = \arg \inf_{\gamma^2 \geq 0} \left| \gamma^2 - (T\hat{V}_{GT} - 1) \right|
\]

with \( \hat{V}_{GT} \) as in (2.4). If \( u = \hat{U}_{TG} \) and \( v = \hat{V}_{GT} \), the function to minimize becomes

\[(1 - u - nuv)\gamma^2 + 1 - u - Nv.\]

We have a proof, which we omit for brevity, that for any possible sample there is a unique solution; we denote such a solution by \( \hat{\gamma}_1^2 \) and the corresponding solution for \( T \) by \( \hat{T}_1 = T_1(\hat{\gamma}_1^2) \). As indicated by the simulations in Section 6, the new estimator seems to perform quite well in most samples, but has an extremely large variance when \( 1 - u - nuv \sim 0 \), which can be seen to happen for every type of population at specific sample sizes; this is presumably the reason for the high variance reported in Esty (1986). Thus the estimator would be very good if one had a method to exclude a priori such a condition, which unfortunately is at present not available. In Section 3 we get a better understanding of the quantities involved in the Chao-Lee estimator.

### 3 Preliminary results on new estimators

(I) Our first result shows that the Chao-Lee estimator emerges as a result of a condition on \( T \), namely the coincidence of two estimators of \( U \).

We start by modifying the add\( -\lambda \) method. Instead of the fictitious \( N+1 \) we apply it to the true, but unknown, number \( T \) of species. Then (2.1) becomes

\[
\hat{p}_i(\lambda) = \frac{m_i + \lambda}{\sum_{i=1}^{T}(m_i + \lambda)} = \frac{m_i + \lambda}{n + T\lambda},
\]

and the estimate of \( U \) changes to \( \hat{U}_\lambda = (T - N)\lambda/T\lambda + n \). Now, we can hope to reconcile the modified “add \( \lambda \)” and the unsmoothed Turing-Good estimators by requiring that they agree on the estimate of \( U \), in other words that \( \hat{U}_\lambda = \hat{U}_{TG} \). Somewhat surprisingly, solving for \( T \) we get

\[
\hat{T}_\lambda = \frac{N + n\hat{U}_{TG}/\lambda}{1 - \hat{U}_{TG}}.
\]

We have obtained
Table 1: Comparison of estimators on four populations whose probabilities range from rather uniform to extremely skewed. 1000 iterations per simulation.

| Estimator | Population | $\gamma^2$ | $n=100$ | Mean | Std | MSE | Mean | Std | MSE | Mean | Std | MSE | Mean | Std | MSE |
|-----------|------------|-------------|---------|-------|-----|-----|-------|-----|-----|-------|-----|-----|-------|-----|-----|
| $T_{Chao}$ | 103 | 18 | 18 | 85 | 15 | 21 | 80 | 17 | 26 | 31 | 17 | 71 |
| $T_{Unif}$ | 101 | 13 | 13 | 80 | 9 | 21 | 68 | 8 | 32 | 19 | 3 | 81 |
| $T_{CL}$ | 104 | 15 | 15 | 85 | 12 | 20 | 82 | 15 | 23 | 27 | 7 | 73 |
| $T_{ACE}$ | 104 | 15 | 15 | 85 | 12 | 20 | 80 | 14 | 24 | 28 | 10 | 72 |
| $T_{JK}$ | 117 | 21 | 27 | 97 | 22 | 22 | 91 | 30 | 32 | 39 | 24 | 68 |
| $T_{1}$ | 115 | 50 | 53 | 98 | 42 | 42 | 125 | 66 | 71 | 37 | 16 | 65 |
| $T_{2}$ | 106 | 15 | 16 | 82 | 10 | 21 | 68 | 8 | 32 | 19 | 3 | 80 |
| $T_{^\hat{}\lambda}$ | 112 | 27 | 29 | 94 | 23 | 24 | 97 | 19 | 20 | 26 | 5 | 74 |
| $T_{^\hat{}\lambda_{ACE}}$ | 112 | 27 | 29 | 94 | 23 | 24 | 95 | 20 | 21 | 30 | 8 | 70 |

Lemma 3.1. The extended “add–$\lambda$” and the Turing-Good estimators of $U$ coincide if and only if $T = T_{CL}(\gamma)$ with $\gamma^2 = 1/\lambda$.

From now on we will assume $\gamma^2 = 1/\lambda$.

(II) The explicit expression of $\hat{U}_\lambda$ in terms of the $p_i$’s given by the extended add–$\lambda$ (3.1) and Lemma 3.1 together show that the estimator of $T$ given in (2.2) can be seen as the solution of a first moment estimate, so that one can hope to derive $\hat{\gamma}^2$, or equivalently $\hat{\lambda}$, from a second moment relation; the form of such a relation can be suggested by the discussion at the end of the previous section and by (2.4), together with the requirement that $\gamma^2 \geq 0$ thus $\lambda \geq 0$. This leads to the system

$$\begin{align*}
\hat{U}_\lambda &= \sum_{k=N+1}^{T} \hat{p}_k(\lambda) = \hat{U}_{TG} \\
\hat{\lambda}_2 &= \arg \inf_{\lambda \geq 0} \left| \sum_{k=1}^{T} \hat{p}_k(\lambda)^2 - \hat{V}_{GT} \right|.
\end{align*}$$

The solutions $\hat{T}_2(\hat{\lambda}_2)$ and $\hat{\lambda}_2$ of the above system, together with $\hat{\gamma}^2 = \hat{\lambda}_2^{-1}$, give new estimators; although this seems to improve the estimation in some

Table 2: Comparison of estimators on four populations with probability distributions following Zipf’s law, i.e., proportional to $1/i + T$ for $i = 1, \ldots, T$; $K = 20, 10, 5, 1$. 1000 iterations per simulation.

| Estimator | Population | $\gamma^2$ | $n=100$ | Mean | Std | MSE | Mean | Std | MSE | Mean | Std | MSE | Mean | Std | MSE |
|-----------|------------|-------------|---------|-------|-----|-----|-------|-----|-----|-------|-----|-----|-------|-----|-----|
| $T_{Chao}$ | 97 | 19 | 20 | 93 | 19 | 20 | 91 | 21 | 23 | 88 | 24 | 27 |
| $T_{Unif}$ | 89 | 11 | 15 | 82 | 10 | 21 | 74 | 9 | 27 | 63 | 8 | 38 |
| $T_{CL}$ | 96 | 16 | 16 | 94 | 17 | 18 | 95 | 19 | 20 | 108 | 27 | 28 |
| $T_{ACE}$ | 96 | 16 | 16 | 94 | 17 | 18 | 92 | 18 | 20 | 89 | 20 | 23 |
| $T_{JK}$ | 112 | 30 | 32 | 108 | 37 | 38 | 106 | 50 | 51 | 99 | 32 | 32 |
| $T_{1}$ | 122 | 61 | 65 | 137 | 77 | 85 | 152 | 83 | 97 | 89 | 56 | 57 |
| $T_{2}$ | 90 | 11 | 15 | 82 | 10 | 21 | 74 | 9 | 27 | 63 | 8 | 38 |
| $T_{^\hat{}\lambda}$ | 111 | 28 | 31 | 112 | 27 | 30 | 112 | 21 | 24 | 99 | 15 | 15 |
| $T_{^\hat{}\lambda_{ACE}}$ | 111 | 28 | 31 | 112 | 27 | 30 | 110 | 23 | 25 | 102 | 20 | 20 |
cases, it does appear to have significant flaws: see tables 1 and 2 in Section 6, in particular the estimation for population 3. We think that the problems come from the too simple form of the function minimized in the system.

4 The Bayesian interpretation

To improve things we recall that the add–λ estimated probabilities are the posterior averages of a Dirichlet prior with a single parameter λ (see Good, 1967, Johnson, 1932, Zabell, 1982). The only novelty here is that we consider the prior and the posterior assuming an unknown number T of species; in other words we consider the Bayesian posterior $\rho_{n,T,\lambda}$ given a sample with species records $m_1,\ldots,m_N$, from a Dirichlet prior with parameter $\lambda \geq 1$ on $Q_T = \{p = (p_1\ldots p_{T-1}) : p_k > 0, \sum_{k=1}^{T-1} p_k \leq 1\}$; we notice only that the condition $\lambda \geq 1$ is necessary for the convergence of the prior. This leads to a reinterpretation of the second moment minimization introduced above in terms of the posterior: one way of achieving this is the substitution of $\hat{p}_k(\lambda)^2$ by $E_{\rho_{n,T,\lambda}}(p_k^2)$ in the above system, which then becomes

\[
\begin{align*}
\hat{U}_{\lambda} &= \sum_{k=N+1}^{T} E_{\rho_{n,T,\lambda}}(p_k) = \hat{U}_{TG} \\
\hat{\lambda} &= \text{arg inf}_{\lambda \geq 1} \left| \hat{V}_{TG} - \sum_{k=1}^{T} E_{\rho_{n,T,\lambda}}(p_k^2) \right|.
\end{align*}
\]

Notice that now we need to take $\lambda \geq 1$ as required by the convergence of the prior. This does not cover all possible populations, since $\gamma^2 = \lambda^{-1}$ can exceed one for peculiar populations, but seems to cover the most natural cases (see Section 6 for a discussion on this point). Our proposed estimation is thus to compute

\[
\hat{\lambda} = \text{arg inf}_{\lambda \geq 1} |f(\lambda)| \quad \text{with} \quad f(\lambda) = \hat{V}_{TG} - \sum_{k=1}^{T} E_{\rho_{n,T,\lambda}}(p_k^2),
\]

and then compute $\hat{T}_{\hat{\lambda}}$ from the first equation of the system, which amounts to taking $\hat{T}_{\hat{\lambda}}$ as in (2.2) with $\lambda^{-1} = \gamma^2$, i.e.,

\[
\hat{T}_{\hat{\lambda}} = \frac{N}{1 - \hat{U}_{TG}} + \frac{n\hat{U}_{TG} - 1}{(1 - \hat{U}_{TG})} \hat{\lambda}^{-1}.
\]

For this to make sense we need to study $f(\lambda)$ and show that for every possible sample there is a unique solution to (4.1) and, finally, that $\hat{T}_{\hat{\lambda}} \geq N$. At the end we will have enough information to produce explicit expressions for $(\hat{\lambda}, \hat{T}_{\hat{\lambda}})$. 
It is a standard calculation using Gamma functions that for every \( \lambda \geq 1, \)
\( 1 \leq i \leq T \) and \( \alpha \in \mathbb{N}, \)
\[
E_{\rho_n,T,\lambda}(p_i^\alpha) = \frac{\Gamma(T\lambda + n)}{\Gamma(T\lambda + n + \alpha)} \frac{\Gamma(m_i + \lambda + \alpha)}{\Gamma(m_i + \lambda)},
\] (4.3)
so that we can get an explicit expression for \( f(\lambda). \) To do so, we let \( u = \hat{U} \) and \( v = \hat{V} \) and express \( f(\lambda) \) as function of \( u \) and \( v; \) also let \( q = vn(n-1) + n = \sum_{j>0} j^2 n_j = \sum_{k=1}^T m_k^2. \) From equation (4.3) we get
\[
f(\lambda) = \hat{V} - \sum_{k=1}^T \left( E_{\rho_n,T,\lambda}(p_k^2) \right)
= v - \sum_{k=1}^T \frac{(m_k + \lambda)(m_k + \lambda + 1)}{(T\lambda + n)(T\lambda + n + 1)}
= v - \frac{\sum m_k^2 + n(2\lambda + 1) + T(\lambda^2 + \lambda)}{(T\lambda + n + 1)(T\lambda + n)}
\] (4.4)
\[= v - \frac{vn(n-1) + n + n(2\lambda + 1) + \lambda(\lambda + 1)N+nu/\lambda}{[n + \lambda N+nu/\lambda + 1][n + \lambda N+nu/\lambda]} \]
The next three lemmas identify some properties satisfied by \( \hat{U} = n_1/n \) and
\[\hat{V} = \hat{V}_{GT} = \sum_{j \geq 1} \frac{j(j-1)n_j}{n(n-1)}\]
whatever the values in the sample.

**Lemma 4.1.** For every sample, \( \hat{U} + \hat{V} \leq 1 \)

**Proof of Lemma 4.1.** Since \( q = \sum_{j>0} j^2 n_j \) and \( n = \sum_{j>0} jn_j \) we have that \( \hat{U} + \hat{V} = n_1/n + q - n/n(n + 1) \leq 1 \) is implied by
\[\text{for } n > 1 \]
\[\sum_{j=1}^N j^2 n_j^2 - n \sum_{j=1}^N j^2 n_j^2 \geq n_1^2 - n_1 \sum_{j=2}^N jn_j - n_1 \sum_{j=2}^N jn_j \geq 0 \]
Lemma 4.2. For every sample, \( N + \hat{U} - 1 - n\hat{U} \geq 0 \)

Proof of Lemma 4.2. By the definition of \( \hat{U} \) we have \( N + \hat{U} - 1 - n\hat{U} = N - n_1 + n_1/n - 1 \), then either \( n = n_1 = N \) and the right-hand side becomes 0, or \( N - n_1 \geq 1 \) and the relation holds.

Lemma 4.3. For every sample, \((\hat{V}nN - \hat{V}N + N - n)n = qN - n^2 \geq 0\)

Proof of Lemma 4.3. Expressing \( q, N, \) and \( n \) as functions of the \( n_j \)'s we get

\[
q = \sum_{j>0} j^2 n_j, \ N = \sum_{j>0} n_j \text{ and } n = \sum_{j>0} jn_j.
\]

Then

\[
qN - n^2 = \left( \sum_{j>0} j^2 n_j \right) \left( \sum_{k>0} n_k \right) - \left( \sum_{j>0} jn_j \right)^2
\]

\[
= \left( \sum_{j>0} j^2 n_j^2 + \sum_{j>0} \sum_{k \neq j} j^2 n_j n_k \right) - \left( \sum_{j>0} j^2 n_j^2 + \sum_{j>0} \sum_{k \neq j} jn_j kn_k \right)
\]

\[
= \sum_{j>0} \sum_{k \neq j} (j^2 - jk) n_j n_k
\]

\[
= \sum_{j>0} \sum_{j<k} (j^2 - jk + k^2 - kj) n_j n_k
\]

\[
= \sum_{j>0} \sum_{j<k} (j - k)^2 n_j n_k \geq 0.
\]

The above properties can be summarized in the following conditions for \( u \) and \( v \), which we assume from now on:

\begin{align*}
0 &\leq u \leq 1 \quad (4.5) \\
0 &\leq v \leq 1 \\
1 &\geq u + v \\
0 &\leq N + u - 1 - nu \quad (4.6) \\
0 &\leq vnN - vN + N - n \quad (4.7)
\end{align*}

We can now give an explicit expression for \( \hat{\lambda} \).

Theorem 4.1. With \( f(\lambda) \) as in equation (4.4) we have

\[
\hat{\lambda} = \arg \inf_{\lambda \geq 1} \left| f(\lambda) \right|
\]

so that

\[
\hat{\lambda} = \begin{cases} 
1, & \text{for } \beta_1 < \lambda_2 \text{ and } 1 \geq \lambda_2; \ i.e., \ \frac{2-v(N+1)}{2-v+vN} \leq u \leq 1-v, \\
\lambda_2, & \text{for } \beta_1 < \lambda_2 \text{ and } \lambda_2 \geq 1; \ i.e., \ 1-Nv < u \leq \frac{2-v(N+1)}{2-v+vN}, \\
\infty, & \text{for } \lambda_2 \leq \beta_1; \ i.e., \ 0 \leq u \leq 1-Nv.
\end{cases}
\]
where \( \beta_1 = -n/N \) is the largest singularity of \( f(\lambda) \) and

\[
\lambda_2 = \frac{1 - u - v + uv - uvn}{Nv + u - 1}.
\]

**Proof of Theorem 4.1.** The equation \( f(\lambda) = 0 \) has solutions:

\[
\lambda_1 = \frac{-2n + nu}{N} \leq 0, \quad \lambda_2 = \frac{1 - u - v + uv - uvn}{Nv + u - 1}
\]

The root \( \lambda_1 \) is always non positive and thus it is not interesting and if \( \lambda_2 \geq 1 \), it achieves the required minimum. To evaluate the other cases note that the function \( f(\lambda) \) has two poles

\[
\beta_1 = -\frac{n - 1}{N}, \quad \beta_2 = -\frac{n - 1}{N} - \frac{1 - u}{N}
\]

and \( \lambda_1 < \beta_2 < \beta_1 \). Moreover,

\[
\lim_{\lambda \to \beta_1^+} f(\lambda) = \infty \cdot \text{sgn} \left( \frac{(u - 1)(vnN - vN + N - n)}{N^2} \right) = -\infty
\]

by (4.5) and (4.7), and

\[
\lim_{\lambda \to +\infty} f(\lambda) = \frac{Nv + u - 1}{N}.
\]

We now verify that

**Lemma 4.4.** \( f(\lambda) \) is increasing for \( \lambda > \beta_1 \).

**Proof of Lemma 4.4.** Let \( g(\lambda) \) be such that

\[
f'(\lambda) = \frac{(1 - u)g(\lambda)}{(n + N\lambda)(1 - u + n + N\lambda)}.
\]

Then

\[
\lim_{\lambda \to \beta_1^+} g(\lambda) = n(1 - u)^2(vnN - vN + N - n) > 0
\]

by (4.7). Note that \( g \) satisfies

\[
g'(|\lambda) = 2N^2(\lambda + u - 1 - nu)\lambda + 2nN(-1 - n + 2N + u - Nu - Nv + nNv + Nuw - nNuv)
\]

with the leading coefficient nonnegative by (4.6). Therefore, if \( \lambda > \beta_1 = -n/N \)

\[
g'(\lambda) > 2nN(1 - u)(vnN - vN + N - n) \geq 0
\]
A new estimator for the number of species

again by (4.7). Thus $g' > 0$ for all $\lambda > \beta_1$ and, by (4.10), $g > 0$ for all $\lambda > \beta_1$ and since the other factors in $f'$ are also positive, we have that $f' > 0$ for all $\lambda > \beta_1$ as required.

Now there are three possibilities.

1. If $u \leq 1 - Nv$ then from (4.9) and Lemma 4.4, it follows that $f < 0$ for all $\lambda > \beta_1$ and increasing, thus
   \[
   \hat{\lambda} = \arg \inf_{\lambda \geq 1} |f(\lambda)| = \arg \max_{\lambda \geq 1} f = +\infty.
   \]

2. If $1 - Nv < u <$ then from (4.8) $\lambda_2 \geq 1$ is equivalent to $u \leq 2 - v(N + 1)/2 - v + vn$, in which case $\hat{\lambda} = \lambda_2$.

3. If $2 - v(N + 1)/2 - v + vn < u$ then $\lambda_2 < 1$ and by Lemma 4.4
   \[
   \hat{\lambda} = \arg \inf_{\lambda \geq 1} |f(\lambda)| = \arg \min_{\lambda \geq 1} f = 1
   \]

The conditions on $u$ and $v$ are translated into those for $\lambda_2$ and $\beta_1$ by direct calculation.

Finally, from (4.2) we get the estimator for $T$:

\[
\hat{T}_\lambda = \frac{N + n\hat{U}_{TG}/\hat{\lambda}}{1 - \hat{U}_{TG}} = \begin{cases}
  n\frac{N + n_1}{n - n_1}, & \text{for } \beta_1 < \lambda_2, \\
  \frac{NN_1}{n - n_1}, & \text{if } \lambda_2 < \beta_1.
\end{cases}
\]

or, alternatively,

\[
\hat{T}_\lambda = \begin{cases}
  \frac{N + nu}{1 - u}, & \text{for } \frac{2 - v(N + 1)}{2 - v + vn} \leq u \leq 1 - v, \\
  \frac{N - Nv - nu}{1 - u - v + uv - wn}, & \text{for } 1 - Nv \leq u \leq \frac{2 - v(N + 1)}{2 - v + vn}, \\
  \frac{N}{1 - u}, & \text{for } 0 \leq u \leq 1 - Nv.
\end{cases}
\]

Notice that when $\hat{\lambda} = \infty$ we get $\hat{T}_\lambda = \hat{T}_{Unif}$.

5 Estimate of species distribution and confidence intervals for $T$

We now use the posterior average probability of each species together with $\hat{T}_\lambda$ and $\hat{\lambda}$ as an estimate of the species probabilities. For the observed species, which is to say $i = 1, \ldots, N$, this amounts to

\[
\hat{\rho}_i = E_{\rho_n, \hat{T}_\lambda, \hat{\lambda}}(y_i) = \frac{m_i + \hat{\lambda}}{\hat{T}_\lambda + n} = \frac{(m_i + \hat{\lambda})(1 - \hat{U}_{TG})}{n + N\hat{\lambda}}
\]

(5.1)
This expression is correct also for $\hat{\lambda} = \infty$, in which case all species are estimated to have probability $(\hat{T})^{-1}$. Also note that these values are close to the unbiased estimator $m_i/n$ of the probability of the $i$-th species and can be seen as a mixture of the Laplace add-$\lambda$ and Turing-Good estimators since they are obtained by adding $\lambda$ to the frequency $m_i$ of the $N$ observed species (recall that $n = \sum_{i=1}^{N} m_i$), but only after assigning the probability $\hat{U}$ to the event that we will observe a new species; the estimate for each of the $N$ species is then reduced by the factor $1 - \hat{U}$ to compensate for this and, in fact,

$$\left(\hat{T}\lambda - N\right) = \frac{\hat{\lambda}(1 - \hat{U}_{TG})}{n + N\hat{\lambda} + \hat{U}_{TG}} = \hat{U}_{TG}.$$

This is likely to be a sensible way to make the attenuation of the Laplace estimator finite (Orlitsky et al., 2003).

For the unobserved species, a simple approach would be to uniformly split the probability $\hat{U}$ among the $\hat{T}\lambda - N$ unobserved species which would give

$$\frac{\hat{U}}{\hat{T}\lambda - N} = \frac{\hat{\lambda}}{\hat{T}\lambda + n} = \frac{\hat{\lambda}(1 - \hat{U})}{n + N\hat{\lambda}}.$$

On the other hand, notice that, since one can rewrite the first equation of the system as $1 - \sum_{k=1}^{N} \hat{p}_k(\lambda) = 1 - \hat{U}_{TG}$, the reconciliation method never used the moments of the $p_i$’s for $i > N$; therefore, we have some freedom in assigning the estimations of the $p_i$’s for $i > N$. These probabilities can, in fact, be estimated by taking into account the meaning of $\lambda^{-1} = \gamma^2$ as normalized species variance; we could then assign probabilities to the unobserved species to achieve the estimated normalized variance $\hat{\gamma}^2$. For simplicity, we actually use a closely related function and focus on $\sum_{k=1}^{N} p_k^2$ and its estimator $\hat{V}$. This is a valid approach except when $u < 1 - N\hat{v}$, in which case $f(\lambda) < 0$ and $\hat{V}$ turns out to be too small to be a reasonable estimate of $\sum_{k=1}^{N} p_k^2$; in that case we replace $\hat{V}$ with $\sum_{k=1}^{N} E_{\rho,\hat{T}\lambda,\hat{\lambda}}(p_k^2)$. Clearly $\sum_{k=1}^{N} (E_{\rho_n,T,\lambda}(p_k))^2 \leq \hat{V} \vee \sum_{k=1}^{N} E_{\rho_n,T,\lambda}(p_k^2)$ by Jensen’s inequality, and thus we require that the estimates $\hat{p}_k$ of the probabilities of the unobserved species satisfy:

$$\sum_{k=N+1}^{\hat{T}\lambda} (\hat{p}_k)^2 = \left(\hat{V} \vee \sum_{k=1}^{\hat{T}\lambda} E_{\rho_n,T,\lambda}(p_k^2) \right) - \sum_{k=1}^{N} \left(E_{\rho_n,T,\lambda}(p_k)\right)^2 =: \hat{V} \quad (5.2)$$

We can use any two-parameter distribution, such as for instance

$$p_i = c\alpha^{-i-N} \text{ for } i = N + 1, \ldots, \hat{T}\lambda, \quad (5.3)$$
and require that
\[
\sum_{i=N+1} \hat{T}_\lambda p_i = \hat{U}_{TG} \quad \text{and} \quad \sum_{i=N+1} \hat{T}_\lambda p_i^2 = \hat{V}.
\]
Solving for \(c\) and \(\alpha\) gives the estimated unobserved probabilities \(\hat{p}_i = p_i(c, \alpha)\). It is easily seen that if \(T > N\), then
\[
\alpha(1 - \alpha) \approx \frac{\hat{U}_{TG}}{\hat{V}} \quad \text{and} \quad c \approx \frac{\hat{U}_{TG}(1 - \alpha)}{\alpha}.
\]
(5.4)

In Section 6 below we generate confidence intervals by resampling based on the estimated species probabilities as obtained in (5.1), (5.3) and (5.4), with \(\hat{V}\) as in (5.2).

6 Simulations and experiments

In this section we present some typical numerical simulations and tests, out of a large number of similar tests, of the performance of several estimators compared to those we have developed here; in particular, besides \(\hat{T}_1\) (see Section 2), \(\hat{T}_2\) (see Section 3 (II)) and \(\hat{T}_\lambda\), we compare the general lower bound \(\hat{T}_{\text{Chao}}\) (Chao, 1984), \(\hat{T}_{\text{Unif}} = \hat{T}_{CL}(0); \hat{T}_{CL} = \hat{T}_{CL}(\hat{\gamma}^2)\) see (2.2); \(\hat{T}_{\text{ACE}}\) (see Chao et al., 1993), which is the ACE method applied to the Chao-Lee estimator; the Jackknife estimator with optimal parameter \(\hat{T}_{JK}\) from Burnahm and Overton (1979); and, finally, \(\hat{T}_{\lambda_{\text{ACE}}}\), which amounts to applying the ACE method to our estimator \(\hat{T}_\lambda\). In table 1 we consider several simulated populations with increasing values of \(\gamma^2\):

1. a population with \(p_i\)'s \(\sim N(1, 0.1)\) which turns out to have \(\gamma^2 \approx 0.009\) and thus it is quite uniform;

2. a population with \(p_i\)'s \(\sim U[0, 1]\) which has \(\gamma^2 \approx 0.3317\), thus less uniform;

3. a population with \(p_i\)'s \(\sim Exp(1)\) and \(\gamma^2 \approx 0.9992\), thus nonuniform; and, finally,

4. an extremely skewed population based on \(p_i\)'s \(\sim \Gamma(0.1, 5)\) which results in \(\gamma^2 \approx 9.1289\).

Each population’s distribution is generated from \(T\) independent and identically distributed random variables, normalized to sum to 1, and then the resulting \(\gamma^2\) is determined as normalized interspecies variance.
In table 1 we report simulations of a simple random sample of size 100 from each population by showing the mean, SD and the mean square error. One can see that in populations (1) – (3) the performance of the estimators depends on $\gamma^2$: in (1), the best estimation (in terms of mean and MSE) is given by $\hat{T}_{Unif}$, in (2) by $\hat{T}_{JK}$, and in (3) by $\hat{T}_{\lambda}$; but that $\hat{T}_{\lambda}$ is performing quite well in each case. Overall, it turns out that this estimator is the one which best adapts to the different values of $\gamma^2$. One can also clearly see that the previous two estimators $\hat{T}_1$ and $\hat{T}_2$ had serious defects. As we mentioned in the introduction, none of the estimators is able to deal with population (4). In general, all estimators would eventually converge but that requires samples of the order of $10^{20}$ because of the smallness of some species probabilities.

In figure 1, to confirm this picture, we simulated sequential simple random samples of increasing sizes (up to $n = 500$). We note that overall $\hat{T}_{\lambda}$ is always close to giving the best results. In figure 1 we plot data of the comparison between $\hat{T}_{\lambda}$ and the Chao-Lee estimator for population (3); this is one of the most evident cases, but a similar improvement happens for many populations, with some exceptions which we briefly indicate further below.

![Figure 1: Comparison of $\hat{T}_{\lambda}$ with $T_{CL}$ and $T_{Chao}$ on a sequential simple random sample from population (3) with $\gamma^2 = 0.9992$, $T = 100$ and $n \in [20, 500]$.](image-url)
Table 3: Summary of confidence interval performances at the given confidence level. Tests are made for populations (1)-(3) and intervals are obtained from the probabilities estimated in Section 5 by resampling. Each entry reports the fraction of hits, and the average interval size in brackets.

| Confidence level | 1               | 2               | 3               |
|------------------|-----------------|-----------------|-----------------|
| 90%              | 93% (1115)      | 92% (821)       | 80% (707)       |
| 95%              | 95% (1225)      | 98% (889)       | 89% (827)       |
| 99%              | 97% (1520)      | 100% (1064)     | 98% (977)       |

In table 2, we repeated the same simulations as in table 1, this time with four distributions following Zipf’s law (from Zipf, 1935, see also Pietronero et al., 2001); in particular, \( p_i = 1/i + K \) for \( i = 1, \ldots, 100 \). We consider \( K = 20, 10, 5, 1 \) which correspond to increasing values of \( \gamma^2 \). Once again, different estimators achieve the best performance (Chao, Chao-Lee twice, and ours); for this kind of data \( \hat{T}_{CL} \) is the overall best estimator, but \( \hat{T}_\lambda \)
assigns acceptable values for all cases. Notice that the best performance occurs for $\gamma^2 \approx 2.6$, which is outside the admissible range. ACE does not really improve the results.

Next, table 3 shows some simulations about confidence intervals for $T$ based on samples of size $n = 400$ computed from $\hat{T}_\lambda$; confidence intervals are obtained by estimating the species probabilities $p_i$’s as described in Section 5 and then resampling 1000 times from the estimated population distribution. This process is repeated 100 times and table 3 indicates, for populations (1)-(3), respectively, the percentage of times the confidence intervals contain the true value of $T = 1000$ and the average size of the confidence interval. The hitting percentage comes out remarkably well, due to the good approximation of the true population distribution by the estimated one.

Next, figure 2 shows a sequential estimation of the initial two-digit distribution based on the populations of counties in the states of the U.S. whose names begin with the letter “A.” There are $T = 90$ possible initial pairs of digits and the distribution is known to follow Benford’s law (Fewster, 2009; Hill, 1995). Sample sizes range from 10 to 184, and for each size repeated random sampling of that size is taken 100 times and averaged. One can see that the two estimators’s performance depends also on the size of the sample; to compare the two we have taken the MSE, over the sample size, from

**Figure 3:** Comparison of $\hat{T}_\lambda$, $T_{CL}$, $\hat{T}_{\lambda_{ACE}}$, $\hat{T}_{ACE}$ and $T_{Chao}$ on a pseudo-Benford’s law with $p_i = \text{const} \log(1 + 1/i)$, $i = 1, \ldots, 100$. Here $\gamma^2 \approx 3.54$ and $n \in [20, 1000]$. 
A new estimator for the number of species

97

Figure 4: Comparison of estimations, empirical and using $\hat{p}_i$, of the two-digit Benford’s law from July 1, 2008 census data of the populations of 184 counties for the States with initial letter “A” taken from U.S Census Bureau. Observed were $N = 67$ out of $T = 90$ classes. Here $\hat{T}_\lambda = 91$.

the true value of $T$, which turned out to be 48 for our estimator and 99 for Chao-Lee. On the other hand, figure 3 studies a pseudo-Benford distribution in which $p_i = \text{const} \log (1 + 1/i)$ for $i = 1, \ldots, 100$, instead of $i$ ranging from 10 to 99. In this case one can see the effect of the ACE method: the Chao-Lee estimator, which is quite inefficient at such high $\gamma^2$, is greatly improved by ACE, while ours, which already has an acceptable performance, is only slightly improved by ACE.

Finally, in figure 4 we use the method of Section 5 to approximate the entire distribution, Benford’s law in this case, improving with respect to plain distribution estimation. We use the entire Census set of $n = 184$ counties. The sample size is such that likely not all pairs were observed. Indeed, we have $N = 67$ observed pairs and empirical estimates of the pair probabilities show a total variation distance from Benford’s law of 0.26. On the other hand, $\hat{T}_\lambda = 91$, the probabilities estimated by $\hat{p}_i$ give a total variation distance of 0.185 and practically each single $p_i$ is better estimated. Notice also that the only species incorrectly predicted to exist (the one numbered 91) is estimated to have a rather small probability, thus partially correcting the prediction error.

Acknowledgement. This work was done during visits by one of us (CCAS) to the Università di Roma, Tor Vergata; Università di Milano-Bicocca; and Università di Firenze. He takes pleasure in thanking those universities for their warm hospitality and GNAMPA for its support. We would also like to
thank J.S. Rao, M. Scarsini, J. Sethuraman and S.R.S. Varadhan for helpful comments.

References

BENFORD, F. (1938). The law of anomalous numbers. *Proc. Am. Philos. Soc.*, 78, 551–572.

BISHOP, Y.M.M., Fienberg, S.E. and HOLLAND, P.W. (1975). *Discrete multivariate analysis: theory and practice*. MIT Press, Cambridge.

BOENDER, C.G.E. and RINNOY KAN, A.H.G (1987). A multinomial Bayesian approach to the estimation of population and vocabulary size. *Biometrika*, 74, 849–856.

BÖHNING, D. and SCHÖN, D. (2005). Nonparametric maximum likelihood estimation of population size based on the counting distribution. *J. R. Stat. Soc. Ser. C Appl. Stat.*, 54, Part 4, 721–737.

BÖHNING, D., SUPPAWATTANABE, B., KUSOLVISITKUL, W. and VIVATWONGKASEM, C. (2004). Estimating the number of drug users in Bangkok 2001: A capture-recapture approach using repeated entries in the list. *Eur. J. Epidemiol.*, 19, 1075–1083.

BROSE, U., MARTINEZ, M.D. and WILLIAMS, R.J. (2003). Estimating species richness: sensitivity to sample coverage and insensitivity to spatial patterns. *Ecology*, 84, 2364–2377.

BUNGE, J. and FITZPATRICK, M. (1993). Estimating the number of species: a review. *J. Amer. Statist. Assoc.*, 88, 364–373.

BURNAHM, K.P. and OVERTON, W.S. (1979). Robust estimation of population size when capture probabilities vary among animals. *Ecology*, 60, 927–936.

BURTON D. (2005). *The history of mathematics: an introduction*. McGraw-Hill.

CHAO, A. (1984). Non-parametric estimation of the number of classes in a population. *Scand. J. Stat.*, 11, 265–270.

CHAO, A. (2004). Species richness estimation. In *Encyclopedia of Statistical Sciences* (N. Balakrishnan, C. B. Read and B. Vidakovic, eds.). Wiley, New York.

CHAO, A. and LEE, S.M. (1992). Estimating the number of classes via sample coverage. *J. Amer. Statist. Assoc.*, 87, 210–217.

CHAO, A., MA, M.-C. and YANG, M.C.K. (1993). Stopping rules and estimation for recapture debugging with unequal failure rates. *Biometrika*, 80, 193–201.

CHAO, A., HWANG, W.-H., CHEN, Y.-C. and KUO, C.-Y. (2000). Estimating the number of shared species in two communities. *Statist. Sinica*, 10, 227–246.

CHURCH, K.W., GALE, W.A. (1991). Enhanced Good-Turing and Cat-Cal: two new methods for estimating probabilities of English bigrams. *Comput. Speech Lang.*, 5, 19–54.

DARROCH, J.N. and RATCLIFF (1980). A Note on Capture-Recapture Estimation. *Biometrics*, 36, 149–153.

EFRON, B. and THISTED, R. (1976). Estimating the number of unseen species: how many words did Shakespeare know? *Biometrika*, 63, 435–467.
esty, W.W. (1985). Estimation of the number of classes in a population and the coverage of a sample. *Mathematical Scientist*, **10**, 41–50.

esty, W.W. (1986). The size of a coverage. *Numismatic Chronicle*, **146**, 185–215.

fewster, R.M. (2009). A simple explanation of Benford’s Law. *Amer. Statist.*, **63**, 26–32.

gandolfi, A. and sastri, C.C.A. (2004). Nonparametric estimations about species not observed in a random sample. *Milan J. Math.*, **72**, 81–105.

good, I.J. (1953). The population frequencies of species and the estimation of population parameters. *Biometrika*, **40**, 237–266.

good, I.J. (1965). *The estimation of probabilities: an essay on modern bayesian method*. Research Monograph No. 30, MIT Press, Cambridge, MA.

good, I.J. (1967). A Bayesian significance test for multinomial distributions. *J. Roy. Statist. Soc. Ser. B*, **29**, 399–431.

good, I.J. and toulin, G. (1956). The number of new species and the increase in population coverage when a sample is increased. *Biometrika*, **43**, 45–63.

harris, B. (1968). Statistical inference in the classical occupancy problem: unbiased estimation of the number of classes. *J. Amer. Statist. Assoc.*, **63**, 837–847.

hill, T.P. (1995). The significant-digit phenomenon. *Amer. Math. Monthly*, **102**, 322–327.

jeffreys, H. (1961). *Theory of probability*. Clarendom Press, Oxford, Third Edition.

johnson, W.E. (1932). Probability: the deductive and inductive problems. *Mind*, **49**, 409–423.

laplace (1995). *Philosophical essays in probabilities*. Springer Verlag, New York.

lewand, R.E. (2008). Relative frequencies of letters in general English plain text. *Cryptographical Mathematics*.

lijd, A.; mena, H.R. and prünster, I. (2007). Bayesian nonparametric estimation of the probability of discovering new species. *Biometrika*, **94**, 769–786.

lindsay, B.G. and roeder, K. (1987). A unified treatment of integer parameter models. *J. Amer. Statist. Assoc.*, **82**, 758–764.

mao, C.X. (2004). Predicting the conditional probability of discovering a new class. *J. Amer. Statist. Assoc.*, **99**, 1108–1118.

marchand, J.P. and schroock, F.E. (1982). On the estimation of the number of equally likely classes in a population. *Comm. Statist. Theory Methods*, **11**, 1139–1146.

orlitsky, A., santhanam, N.P. and Zhang, J. (2003). Always good Turing: asimptotically optimal probability estimation. *Science*, **302**, 427–431.

pietrone, L., tosatti, E., tosatti, V. and vesprini, A. (2001). Explaining the uneven distribution of numbers in nature: The laws of Benford and Zipf. *Phys. A*, **293**, 297–304.

shen, T-J., chao, A. and lin, C-F. (2003). Predicting the number of new species in further taxonomic sampling. *Ecology*, **84**, 798–804.

tao, T. (2009). Benford’s law, Zipf’s law, and the Pareto distribution, Terence Tao’s blog. [http://terrytao.wordpress.com/2009/07/03/benfords-law-zipfs-law-and-the-pareto-distribution/](http://terrytao.wordpress.com/2009/07/03/benfords-law-zipfs-law-and-the-pareto-distribution/)
L. Cecconi, A. Gandolfi and C.C.A. Sastri

zabell, s.l. (1982). W. E Johnson’s “Sufficientness” postulate. Ann. Statist., 10, 1090–1099.

zipf, g.k. (1935). The psychobiology of language; an introduction to dynamic philology. Houghton Mifflin, Boston.

Lorenzo Cecconi and Alberto Gandolfi
Dipartimento di Matematica
U. Dini, Università di Firenze
Viale Morgagni 67/A
50134 Firenze, Italy
E-mail: cecconi@math.unifi.it; gandolfi@math.unifi.it

Chelluri C. A. Sastri
Missouri University of Science and Technology, Rolla
Missouri, USA
and Department of Mathematics and Statistics, Dalhousie University
Halifax, NS B3H 3J5, Canada
E-mail: sastri@mathstat.dal.ca

Paper received: 16 December 2009; revised: 21 November 2010.