# ESTIMATION OF THE NUMBER OF SPECIES FROM A RANDOM SAMPLE

Alberto Gandolfi
Dipartimento di Matematica U. Dini,
Università di Firenze,
Viale Morgagni 67/A, 50134 Firenze, Italy
email: gandolfi@math.unifi.it

and

C.C.A. Sastri
Department of Mathematics and Statistics
Dalhousie University, Halifax, Nova Scotia Canada B3H 3J5
email: sastri@mathstat.dal.ca

Dedicated to the memory of Thyagaraju Chelluri, a wonderful human being who would have become a fine mathematician had his life not been cut tragically short.

ABSTRACT. We consider the classical problem of estimating $T$, the total number of species in a population, from repeated counts in a simple random sample and propose a new algorithm for treating it. In order to produce an estimator $\hat{T}$ we actually start from the estimation of a related quantity, the unobserved probability $U$. In fact, we first show that an estimation of $T$ can be obtained by requiring compatibility between the Laplace add-one (or add-$\lambda$) estimator and the Turing-Good estimator $\hat{U}_{TG}$ of $U$; the estimators obtained in this way concide with those of Chao-Lee and of Horvitz-Thompson, depending on $\lambda$. On the other hand, since the Laplace formula can be derived as the mean of a Bayesian posterior with a uniform (or Dirichlet) prior, we later modify the structure of the likelihood and, by requiring the compatibility of the new posterior with $\hat{U}_{TG}$, determine a modified Bayesian estimator $\hat{T}'$. The form of $\hat{T}'$ can be again related to that of Chao-Lee , but provides a better justified term for their estimated variance. $\hat{T}'$ appears to be extremely effective in estimating $T$, for instance improving upon all existing estimators for the standard fully explicit Carothers data. In addition, we can derive estimations of the population distribution, confidence intervals for $U$ and confidence intervals for $T$; these last appear to be the first in the literature not based on resampling.

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

## 1. Introduction

We consider the classical problem of estimating the number $T$ of species in a population, and, subsequentely, 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. Problems of this kind arise in a variety of settings: for example, when sampling fish from a lake or insects in a forest (see, for instance, Shen et al (2003) on how to use estimates of $T$ to predict further sampling, or Brose et al (2003)); or when estimating the size of a particular population (see Böhning et al (2004)); or when trying to guess how many letters an alphabet or how many specific groups of words a language contains (see Church and Gale (2006)) or how many words a writer knows (see Efron and Thisted (1976)); or, even, when determining how many different coins were coined by an ancient population Esty (1986)). Because of its great interest this has become a classic in probability and there has been a great number of studies suggesting methods for the estimation of $T$. See, for instance, Bunge and Fitzpatrick (1993) for a review through 1993 and Gandolfi and Sastri (2004) for some further details.

A quantity closely related to $T$ has also been studied intensively. Each of the species which have not been observed in the sample has some probability of being selected next if we were to continue sampling, and the sum $U$ of all these probabilities is the probability that the next trial will result in the appearance of an unobserved species. The estimation of $U$ is also of interest in a number of situations, for instance when deciding whether to carry a special container for possible new species or whether to reserve part of a code for unusual words; it is also of interest in dealing with genomic datasets while evaluating the probability of discovering new genes by sequencing additional sequences of DNA fragments (see Mao (2004) and Lijoi et al (2007)) or, in general, in finding stopping rules. We can turn this second question into a simplified version of our original problem by assuming that there are $N+1$ species, the $N$ observed ones and the "new" species with probability $U$; the main issue becomes then the estimation of the probabilities of the various species and especially for the new one. For this and other reasons that we shall see, the estimations of $T$ and $U$ are closely intertwined.

We now fix some notation before proceeding. Assume that the population from which the sample is drawn has a total of $T$ species (which we sometimes will call states) having proportions $p_1, p_2, \cdots, p_T.$; and that in a sample $x_1, x_2, \cdots, x_n$ of size $n$ there are $N$ observed species. For $i = 1, \cdots, T$, let $m_i$ be the number of observations of the species $i$ in the sample, so that $\sum_{i=1}^{N} m_i = n$. For later purposes, let $R$ denote the number of species observed

more than once and assume that the $m_i$'s are given one of the possible orders in which $m_1, \ldots, m_R \geq 2$, $m_{R+1}, \ldots, m_N = 1$ and $m_i = 0$ for $i = N+1, \ldots, T$. Also, for $j = 1, \cdots, n$, 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 $U = \sum_{i:L_n(i)=0} p_i$ is the unobserved probability.

The first attempt to estimate $U$ can be extracted from Laplace (see Laplace (1995) and Orlitsky et al (2003)) who suggested an "add-one" estimator: this consists in adding one to the number of observations of each species plus an additional one for the "unobserved" species. The estimate of the unobserved probability becomes: $\hat{U}_L = \frac{1}{1+\sum_{i\geq 0}(m_i+1)} = \frac{1}{1+n+N}$. Laplace's method provides also an estimate of the probability of each observed species $i$ as $\hat{p}_i = \frac{m_i+1}{1+n+N}$.

If adding "one" does not sound like a sensible choice, then one can easily develop an "add $\lambda$" estimator $\hat{U}_{L,\lambda}$, in which some positive value $\lambda$ is added to each species' frequencies (including the unobserved one). To see how it works, just change the "1"'s into "$\lambda$"'s in the above formulas. A recent advance in the direction of estimating the unobserved probability $U$ appears in Orlitsky et al (2003), in which a quantity is introduced, called attenuation, that measures how well the estimation of $U$ works as the sample gets larger, and in which asymptotically very good estimators are determined.

With a seemingly completely different method, Turing and Good (see Good (1953)) proposed another estimator of $U$. Recall that $n_1$ is the number of species observed exactly once and $n$ the size of the sample; then the Turing-Good estimator for $U$ is some minor modification of:

$$\hat{U} = \frac{n_1}{n}.$$

A plausible rationale for this estimator is that while for species observed at least twice the empirical frequency is already becoming stable and very likely close to the probability of the species, species observed only once are likely to be randomly selected representatives of the collection of the yet unobserved species. In more mathematical terms, Good (1953) has a derivation for the estimation of the probability of the species observed $j$ times. The Turing-Good estimator for the total probability $C_j = \sum_{i:L_n(i)=j} p_i$ of the species observed $j$ times is thus

$$\hat{C}_0 = \hat{U}_{TG} = \frac{n_1'}{n'}$$

for $j = 0$ and

$$\hat{C}_j = \frac{j+1}{n'} n_{j+1}'$$

for $j \geq 1$, where the $n'_j$'s are "smoothed" values of the $n_j$'s and $n' = \hat{C}_0 + \sum_{j>0:n'_{j+1}>0} \hat{C}_j$. Smoothing is a minor modification of the original value and is needed for various reasons, one of which is avoiding the possibility that some observed species are estimated to have zero probability (see Good (1953) and Orlitsky et al (2003)). We adopt here a smoothing which is very close to one also suggested by Good, obtained by letting $n'_j = \max(n_j, 1)$ for $j \geq 2$, so that we use

$$\hat{U}_{TG} = \frac{n_1}{n_1 + \sum_{j>0:n_j>0} \max(n_{j+1}, 1)}.$$

We take $n'_1 = n_1$, and not equal to $\max(n_1, 1)$, because the behaviour of the estimators that we will produce in connection with $U_{TG}$ is better described if we allow the possibility that the unobserved probability is estimated to be zero. Our method of estimation of $T$ will make use of an estimation of $U$, and we choose the smoothed Turing Good estimator.

Other methods for estimating $U$ have been developed, and in particular we refer to Lijoi et al (2007) for a Bayesian method based on the general class of Gibbs-type priors (see also Pitman (2005) and the other references in Lijoi et al (2007) for the definition and properties of such priors). This class contains several known families of priors as particular cases and each such family is based on one or more parameters, which need to be further estimated. In Lijoi et al (2007), for instance, a maximum likelihood estimator is used. It is conceivable that within this wide class some extention of the methods we present here could produce even better results than those we obtained. However, we focus on the Turing-Good estimator since it is more direct and simpler, while still allowing us to achieve very satisfactory results.

Let us describe, at this point, the type of data to which our analysis applies. There are several types of data in which there are classes and associated natural numbers and in which it is typically recorded how many classes are associated to a given natural number. Consider three examples: in sampling from a lake the classes are the different fish species and the associated natural number is the the number of times a species was captured; in surveying the number of occupants of a car, each car is a class and the number of occupants is the associated natural number; in recording the daily number of deaths within a certain population, each day is a class and the number of deaths is the associated natural number.

These examples are substantially different. In the case of car occupants, 0 cannot occur, while it can in the other two cases. In the other two examples, on the other hand, if $k$ classes are associated with the natural number $m$ in a sample of size $n$, say, we are lead to different conclusions. In the number of deaths, we are lead to conclude that each class has about probability $k/N$ of being associated to $m$; while in the fish sampling, we are lead to conclude that there are $k$ classes having probability about $m/n$ of being sampled (which is in accordance, with the exception of $m = 1$, with the above mentioned rationale for the

Turing-Good estimator). The first type of data could be called "Poissonian data", this last could be called "Turing-Good type" data, while the data from the car occupancy survey could be called $zero-truncated$ data (in the specific case of Poissonian type).

Notice, in particular, that in the Poissonian type the natural number 0 plays a different role from that in the Turing-Good type data. In the first, in fact, if $k$ classes were associated to 0, the above procedure would assign a nonzero probability to 0, and thus there are no logical hindrances to observing the natural number 0 also; however, in the Turing-Good type data, the above procedure would lead to a probability 0 of those classes, so that, having probability 0, it is impossible, on a logical basis, that those classes, and thus the natural number 0, are observed: a fact well expressed in the tautology that one cannot observe in a sample that some species were not observed in the sample.

We are interested in the non zero-truncated $Turing-Good$ type data and will thus develope methods and discuss examples for this type of data. Notice, for intance, that in Böhning and Schön (2005) the two types of data are mixed together and estimation methods are applied to all of them. Our estimation method applies only to the two Turing-Good type data reported there, namely the illegal immigrants in the Netherlands on page 724 and the 1995 North American breeding bird abundance survey taken from Norris III and Pollock (1998) on page 735. All other examples reported in the paper are of the Poissonian type, including all those with an explicit value of the number of classes associated to 0.

This is no surprise. Due to the intrinsic impossibility of observing the classes associated to 0, gaining complete knowledge of the number of classes associated to 0 in the Turing-Good type data must require a substantially different process from that used in the sampling, so that it is unusual to have data reporting both the sampling procedure and the complete information about all classes. Two notable exceptions are the data in Carothers (1973) and in Edwards and Eberhardt (1967), resulting from experiments explicitly devoted to the generation of this type of information. We briefly recall the content of the data in section 5 and then test our estimators on them.

It is time to get back to the estimation of $T$. In this direction there are several parametric methods based on assuming some structure of the species distribution; for instance, an estimator devised for the uniform case, in which the probabilities of all species are assumed to be the same is the Horvitz-Thompson

$$\hat{T}_{HT} = \frac{N}{1-U},$$

(see Lindsay and Roeder (1987) and Bishop et al (1975)) and then $U$ can be further estimated, for instance by the Turing-Good method, to get

$$\hat{T}_{HTTG} = \frac{N}{1 - \hat{U}_{TG}}.$$

See Darroch and Ratcliff (1980) and Böhning and Schön (2005). Another estimator developed for the uniform case is a Bayesian estimator (see Marchand and Schroeck (1982)) based on the improper prior on $T$ uniform over the positive integers. Also the method in Böhning and Schön (2005), which is more appropriate for and mostly applied to Poissonian type data, relies on some uniformity assumption, since it assumes that each class in the population has the same (Poisson or Poisson mixture) probability of being associated to a certain natural number.

On the other hand, we want to focus here on non-parametric estimation.

If one, in fact, has no reasonable guess for the form of the distribution then a non-parametric approach is needed. In this direction, Harris (1968), Chao (1984) and Chao and Lee (1992) have proposed some such estimators, of which the most reliable ones seem to be those proposed in Chao and Lee (1992). In our notation these amount to

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

with $\hat{\gamma}^2$ an estimate - for which Chao & Lee make two proposals - of the coefficient of variation of the $p_i$'s. The $\hat{\gamma}^2$'s, however, are determined by somewhat involved procedures and are not fully justified from a theoretical point of view.

We start our work by proposing a comparison between two of the above estimators of $U$: Laplace's "add one" and Turing-Good. In fact, it would make sense to apply Laplace's estimator by adding one to each of the frequencies of all the $T$ species, not just of the arbitrary $N + 1$. Of course, we do not know the true value of $T$, but for any given value the "add one" estimator would estimate a probability of $\frac{m_i+1}{T+n}$ for a species observed $m_i$ times (with $m_i$ possibly equal to 0). Now, we can hope to reconcile the Laplace and Turing-Good estimators by requiring that they assign the same value to $\hat{U}$. Since in the "add one" performed on $T$ species the total probability for the $T - N$ unobserved species is then estimated to be $\frac{T-N}{T+n}$, reconciling the two estimators would imply a value of $T$ which solves $\frac{T-N}{T+n} = \hat{U}_{TG}$. This equation happens to have a single root, which is also larger than or equal to $N$ and thus can serve as an estimator of $T$: $\hat{T}(1) = \frac{N}{1-\hat{U}_{TG}} + \frac{n\hat{U}_{TG}}{1-\hat{U}_{TG}}$. Quite surprisingly, this turns out to be the Chao-Lee estimator with the estimated variance $\hat{\gamma}^2$ equal to 1. This is already something, but it is not such a great achievement since the Chao-Lee estimator with variance 1 is not

so good: Chao and Lee discuss a few cases in which it might make sense, but its inadequacy was the main reason for introducing the estimated variance term; the inadequacy of $\hat{T}$ can also be seen in our table 1 below in which several estimators are applied to the Carothers data (see Carothers (1993) for fully detailed and published data of sampling from a known taxi cab population): the rms error of $\hat{T}$ from the true value is much larger than for most other proposed estimators in the literature. However, the reconciling procedure seems to have to produce a somewhat more meaningful result, so we proceed further.

An estimator with an additional parameter that could be suitably tuned might then be obtained by reconciling Turing-Good with the "add $\lambda$" estimation of $U$. In the above terms we need to solve $\frac{(T-N)\lambda}{T\lambda+n} = \hat{U}_{TG}$, which gives

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

This is nothing else than the Chao-Lee estimator with $\gamma^2 = 1/\lambda$. In this way, we have gone one step forward, and we produced indeed a more flexible estimator, completely reproducing the Chao-Lee result; from the point of view of the estimation of $T$, however, the problem has unfortunately just shifted from estimating $T$ to estimating $\lambda$ or, in fact, $\gamma^2$. At this point, one can clearly resort to methods proposed in the literature on how to estimate either $\lambda$ (see Huand (2006) or Good (1967)) or $\gamma$, or proceed further with the reconciliation.

To pursue the second direction, we really need to understand more about the "add one" and "add $\lambda$" estimators. It turns out, as was probably known already to Laplace, that the probability estimation according to the "add one" method is nothing else but the average species probability under the Bayesian posterior given the sample with a prior uniform over all possible probability distributions on $T$ species. To be more specific, let

$$\Sigma_T = \{p = (p_1, p_2, \cdots, p_T), p_i \geq 0, \sum_{i=1}^{T} p_i = 1\}$$

and consider the uniform prior $\rho$ on $\Sigma_T$. Then, given a sample $x$ the likelihood is

$$\mu(x) = \prod_{j=1}^{n} p_{x_j} = \prod_{i=1}^{T} p_i^{m_i} = \prod_{i=1}^{N} p_i^{m_i}$$

and the posterior becomes

$$\begin{aligned} \rho_{n,T}(d\mu) &= \frac{\mu(x)\rho_{0,T}(d\mu)}{\int_{\Sigma_T}\mu(x)\rho_{0,T}(d\mu)} \\ &= \frac{1}{Z}\mathbf{1}_{\Sigma_T} p_1^{m_1} p_2^{m_2}\cdots p_N^{m_N} \quad dp_1\cdots dp_N dp_{N+1}\ldots dp_T \end{aligned} \tag{2}$$

where $Z = \int_{\Sigma_T} p_1^{m_1} \cdots p_N^{m_N} dp_1 \cdots dp_T$ (note that the constant terms have been cancelled).

We then get the "add one" probability estimation by taking the average species probability under such posterior:

$$E_{\rho_{n,T}}(p_i) = \frac{m_i + 1}{T + n}, \qquad i = 1, \ldots, N \tag{3}$$

$$E_{\rho_{n,T}}(p_i) = \frac{1}{T + n}, \qquad i = N + 1, \ldots, T. \tag{4}$$

In addition, Johnson proposed the use of the broader class of Dirichlet distributions as priors: see Johnson (1932) for the original introduction, Jeffreys (1961) and Good (1965) for various discussions, and Zabell (1982) for a historical description. The Dirichlet distributions depend on one parameter (it is possible to introduce one parameter for each state, but we restrict ourselves to a constant choice) that we here indicate by $\lambda$. The prior $\rho_{0,T,\lambda}$ has then density $c \prod_{i=1}^{T} p_i^{\lambda - 1}$ for some constant $c$ and the posterior becomes

$$\rho_{n,T,\lambda}(d\mu) = \frac{1}{Z_\Lambda} \mathbf{1}_{\Sigma_T} \prod_{i=1}^{T} p_i^{m_i + \lambda - 1} dp_1 \ldots dp_T.$$

As the reader has guessed by now, the average under the posterior starting from the Dirichlet prior distribution becomes the estimated probability using the "add $\lambda$" estimation:

$$E_{\rho_{n,T,\lambda}}(y_i) = \frac{m_i + \lambda}{T\lambda + n}, \qquad i = 1, \ldots, N \tag{5}$$

$$E_{\rho_{n,T,\lambda}}(y_i) = \frac{\lambda}{T\lambda + n}, \qquad i = N + 1, \ldots, T, \tag{6}$$

from which the full Chao-Lee estimator has been previously derived.

The reconcilation between several estimators has thus led us to a Bayesian approach and we now explore in that direction. Besides the method for uniform species distributions mentioned in Marchand and Schroeck (1982), a general Bayesian approach is presented in

Boender and Rinnoy Kan (1987), by starting from a prior distribution of $T$ and, conditionally to $T$, a uniform or Dirichlet($\lambda$) prior on the species probability. This method, however, is seen to depend heavily on the choice of $\lambda$ and thus leads to introducing a (level III) prior on $\lambda$ itself (as suggested in Good (1967)) which in turn requires the introduction of a further parameter (Boender and Rinnoy Kan (1987), formulae (10) and (11)), with then no analytical expression for the posteriors. In the end, this direction seems to include several undetermined choices (the prior on $T$ and the extra parameter at level III) and no simple analytical expression of the estimators.

On the other hand, we are now in a position to improve the reconciliation method. The standard Bayesian posterior and, thus, the "add" estimators do not really reflect the rationale beyond the Turing-Good method, because they treat the species observed once the same as those observed more times. The idea beyond Turing-Good is that, instead, the species observed once and those not observed should be lumped in a single group observed, thus, $n_1$ times. This suggests that a more appropriate likelihood could take into account the fact that only $R = N - n_1$ have been observed more than once and thus give

$$\mu(x) = \prod_{i=1}^{R} p_i^{m_i} (1 - p_1 - \cdots - p_R)^{n_1}.$$

A slightly less standard calculation, carried out in section 2 below, shows that now the average posterior probability with a uniform prior is

$$E_{\rho'_{n,T}}(y_i) = \frac{m_i + 1}{T + n}, \qquad i = 1, \ldots, R \tag{7}$$

$$E_{\rho'_{n,T}}(y_i) = \frac{n_1}{(T - R)(T + n)} + \frac{1}{T + n}, \qquad i = R + 1, \ldots, T. \tag{8}$$

This amounts to an "add one" estimator, with the species observed less than twice sharing the observed frequency $n_1$.

The average value of $U$ under the posterior is just $T - N$ times the last expression and thus reconciling such an estimation with Turing-Good leads to solving the equation $E_{\rho'_{n,T}}(U) = (T - N)(\frac{n_1}{(n_1+T-N)(T+n)} + \frac{1}{T+n}) = \hat{U}_{TG} = \hat{U}$. The only solution in $[N, \infty]$ of such an equation

turns out to have the form

$$\hat{T'} = \frac{(N-n_1)(2-\hat{U}) + n\hat{U} + \sqrt{4(n_1)^2(1-\hat{U}) + (\hat{U})^2(n+N-n_1)^2}}{2(1-\hat{U})}$$

$$= \frac{N}{(1-\hat{U})} + \frac{n\hat{U}}{(1-\hat{U})}\gamma^2 \quad (9)$$

with

$$\gamma^2 = \frac{(n-N+n_1)\hat{U} - 2n_1 + \sqrt{4(n_1)^2(1-\hat{U}) + (\hat{U})^2(n+N-n_1)^2}}{2n\hat{U}}. \quad (10)$$

It also turns out that $1 \geq \gamma^2 \geq 0$ and $\gamma^2 = 0$ iff all states have been observed exactly once (which is to say, $n_1 = N = n$). Thus, we get again the Chao-Lee estimator, but this time with an explicit expression for the $\gamma^2$ term; the expression we get behaves like an interspecies variance, and it does so even more than the values, occasionally exceeding one, suggested by Chao and Lee. The value $\hat{T'}$ thus stands a better chance of being a good estimator of $T$. And indeed our table 1 shows that on the Carothers data it performs much better than $\hat{T}$ and for far from uniform distributions (such as are those labelled $A\gamma$ and $B\gamma$) is even the best estimator available.

It would be possible at this point to start from Jeffrey's distribution. However, there seems to be no clearcut gain in doing so. With Carothers' data, the value of $\lambda$ which would return the true population size with the modified Bayesian estimator is often close to 1 with no easily identifiable patterns in the deviations. It is still conceivable that different contexts require different values of $\lambda$ as suggested in Boender and Rinnoy Kan (1987), but we do not pursue this direction in the present paper and we restrict ourselves to the analysis of $\hat{T}$ and $\hat{T'}$.

Once we have an estimated value of $T$, we can take the average probability under the (modified) Bayesian posterior distribution, and this provides an estimation for the species distribution. Such an estimation problem is very relevant in many contexts, and, by itself, our estimation method produces a new and original estimator. For a direct application and a quick reference to existing methods see Jedynak et al (2005), in which it is also shown how to use the Turing Good and the "add one" estimators to estimate the species probabilities (see section 1 in Jedynak et al (2005)); in the paper, however, the relation between these two estimators is not realized and when the population size is needed (in section 4) it is estimated with an "ad hoc" method.

In addition to what we have discussed so far, we can bootstrap our method to provide an estimation of the distribution of $U$. This is achieved by assuming for $T$ one of the estimated values and by defining

$$P(U \geq \epsilon) = \rho_{n,T}\left(\mu = (p_1, \ldots, p_T) : \mu(U) = \sum_{i:L_n(i)=0} p_i \geq \epsilon\right).$$

If we replace $T$ by $\hat{T}$ then the r.h.s. becomes a function of the sample only, and thus it can provide an estimator for the distribution of $U$:

$$\hat{P}(U \geq \epsilon) = \rho_{n,\hat{T}}\left(\mu = (p_1, \ldots, p_{\hat{T}}) : \mu(U) = \sum_{i:L_n(i)=0} p_i \geq \epsilon\right).$$

Depending on which expression is taken for $\hat{T}$ we get different estimators for the distribution of $U$.

In Almudevar et al (2000) also there is an estimate of the distribution of $U$, but by the way it is built, it relies on the fact that the sample gives a good approximation of the population distribution, a situation which occurs when almost all species have been observed. This amounts to developing an estimator of the "small" residual probability that, even if we think that most if not all species have been observed, still some species with small probabilies have been missed. It is a complementary range of application with respect to ours, as we implement a large correction due to the likely presence of several unobserved species carrying a substantial total probability. This suggests an alternative use of the two estimators of the distribution of $U$ for "large" and "small" samples; the formulation of our estimator suggests,in turn, the use of the positivity of $n_1$ to discriminate between the two. Actually, this could be quite a general argument for statistical tests: when the range of the possible observations is not known and some indicator like $n_1$ is strictly positive, all estimators need to be corrected to take care of the presence of some unobserved states; otherwise, one can use the usual estimators.

Finally, we carry out a second bootstraping. We observe that, for a fixed level $\alpha$, the estimated distribution of $U$ allows us to find real intervals $I$ such that $P(U \in I) \geq \alpha$; such an interval, which can be chosen to contain the Turing-Good estimator of $U$, can be taken as a confidence interval for $U$. Furthermore, observing that $E_{\rho'_{n,T}}(U)$ is a continuous strictly increasing function of $T$, one can take the inverse image of $I$ under $E_{\rho'_{n,T}}(U)$, and interpret this as an $\alpha$-confindence interval for $T$. Thus, this method generates confidence intervals based, and possibly centered, on the estimates $\hat{T}$ and $\hat{T}'$. Calculations for the centered version are carried out in Section 4. The confidence intervals that we provide are the first to be defined without necessity of generating external data: the methods used so far, for

instance in Chao (1984) or Colwell (2006), follow Efron (1981) and require the construction of a "pseudo-population" with random draws from it.

In Section 5 we compute our confidence intervals for some data from Carothers (1973) and Edwards and Eberhardt (1967). Unfortunately, we can make explicit evaluations only for the regular Bayesian, which does not provide an acceptable confidence interval. The exact formula for the modified Bayesian method is computationally too heavy and could not be easily computed even for the Carothers data. Also, asymptotic analysis (see, for instance, Lehmann (1983), sec. 2.6 and 6.7) does not seem suitable to approximate the distribution of $U$; standard calculations show that the regular Bayesian estimate of the distribution of $U$ is, for $n$ large, asymptotically normal with mean $\hat{U}$ and SD $= \frac{\sqrt{\hat{U}(1-\hat{U})}}{\sqrt{n}}$, independent of the initial distribution on the $p_i$'s. Unfortunately, this turns out to be not such a good approximation of our estimate of the distribution of $U$, at least for several small examples and for the Carothers data. A plausible explanation is that the sample size is not large enough for the asymptotics to take place. On the other hand, since in our problem large sample sizes would yield a delta at 0 as estimate of $U$, it is conceivable that standard asymptotic analysis is never applicable to our problem.

On the other hand, calculations with very small size examples show that the standard deviation of the distribution of $U$ computed via the modified Bayesian is a constant multiple times the standard deviation obtained via the regular Bayesian; so we implement a mixed formula in which the range of the interval $I$ estimating $U$ is taken from the regular Bayesian formula, then it is multiplied by a suitable factor and mapped into the $T$ space by the modified Bayesian estimator. The confidence intervals produced by this method turn out to be quite narrow and, nonetheless, cover the true value with a frequency quite close to the level of confidence.

We must remark, though, that the method of calculation, in particular the choice of the above mentioned multiplicative factor, and the coverage of the true value by the confidence interval are not entirely satisfactory, and thus approximating formulas for the modified Bayesian expression and modified definitions of the confidence intervals will be the subject of further research.

## 2. Point estimators of $T$ and of species distribution.

As mentioned above, we consider a population with $T$ species having proportions $p_1, p_2, \cdots, p_T$ and adopt the notation introduced in the previous section.

Our first proposition summarizes the remark that an estimate of $T$ is uniquely determined by reconciling the add$-\lambda$ and the Turing-Good estimators.

**Proposition 2.1.**

Consider a simple random sample of size $n$ drawn with replacement from a population with $T$ species. The only value $\hat{T}(\lambda)$ of $T$ such that both the add$-\lambda$ and the Turing-Good estimators assign the same probability to the collection of unobserved species is

$$\hat{T}(\lambda) = \frac{N}{1-\hat{U}_{TG}} + \frac{n\hat{U}_{TG}}{(1-\hat{U}_{TG})}\frac{1}{\lambda}. \tag{11}$$

PROOF The add$-\lambda$ estimator assigns probability $\frac{m_i+\lambda}{T\lambda+n}$, $i = 1, \ldots, N$ to the observed species and $\frac{\lambda}{T\lambda+n}$, $i = N+1, \ldots, T$ to the unobserved ones; and thus it assigns probability $\frac{(T-N)\lambda}{T\lambda+n}$, to the collection of unobserved species. Let us denote by $\hat{U}_{TG}$ the value of a Turing-Good estimator (for some choice of smoothed constants) of the unobserved probability. By equating the two values $\frac{(T-N)\lambda}{T\lambda+n} = \hat{U}_{TG}$ and solving for $T$ we obtain the result.

□

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

Our aim is now to recall the derivation of the add$-\lambda$ estimator of species probabilities from a Bayesian scheme.

Given $T$, a prior knowledge about the population distribution can be described by a measure $\rho_0 = \rho_{0,T,\lambda}$ on $\Sigma_T$, which we can initially take to be uniform ($\lambda = 1$) or, more generally, Dirichlet($\lambda$) with density $c\prod_{i=1}^{T} p_i^{\lambda-1}$. The classical Bayesian likelihood is $\mu(x) = \prod_{j=1}^{n} p_{x_j} = \prod_{i=1}^{T} p_i^{m_i}$ and the posterior density then becomes $\rho_{n,T,\lambda}(d\mu) = \mu(x)\rho_{0,T,\lambda}(d\mu) = \frac{1}{Z}\prod_{i=1}^{T} p_i^{m_i+\lambda-1} dp_1 \ldots dp_T$.

Posterior densities are easily computed by means of the the projection on

$$Q_T = \{y = (y_1, y_2, \cdots, y_{T-1}), y_i \geq 0, \sum_{i=1}^{T-1} y_i \leq 1\},$$

which is given by

$$\mu(y) = \prod_{i=1}^{T-1} y_i^{m_1}(1-y_1-\ldots-y_{T-1})^{m_T} = \prod_{i=1}^{T} y_i^{m_i}, \tag{12}$$

where the last equality makes sense if we additionally define $y_T = (1-y_1-\cdots-y_{T-1})$. Then the Bayesian posterior becomes

$$\begin{aligned}
&\rho_{n,T,\lambda}(d\mu) \\
&\quad = \frac{\mu(x)\rho_{0,T,\lambda}(d\mu)}{\int_{Q_T}\mu(x)\rho_{0,T}(d\mu)} \\
&\quad = \frac{1}{Z}\mathbf{1}_{Q_T} y_1^{m_1+\lambda-1} y_2^{m_2+\lambda-1}\ldots y_{T-1}^{m_{T-1}+\lambda-1}(1-y_1-\cdots-y_{T-1})^{m_T+\lambda-1} \quad dy_1\ldots dy_{T-1}
\end{aligned}$$

where $Z = \int_{Q_T} y_1^{m_1+\lambda-1} y_2^{m_2+\lambda-1}\ldots y_{T-1}^{m_{T-1}+\lambda-1}(1-y_1-\cdots-y_{T-1})^{m_T+\lambda-1}dy_1\ldots dy_{T-1}$.

As mentioned in the introduction, we look for a slightly modified version of the likelihood obtained by lumping together the states that have been observed zero or one times, since this actually corresponds more directly to the rationale behind the Turing-Good estimator. As also mentioned, we consider only $\lambda = 1$ and, given that the number of states observed more than once is $R = N - n_1$, we take

$$\mu'(x) = \prod_{i=1}^{R} p_i^{m_i}(1-p_1\cdots-p_R)^{m_{R+1}+\cdots+m_N}$$

(where if $R = T$ we take $m_{R+1}+\cdots+m_N = 0$ and $0^0 = 1$) with its projected version,

$$\mu'(y) = \prod_{i=1}^{R} y_i^{m_i}(1-y_1\cdots-y_R)^{n_1}$$

(taken with $y_T = 1-y_1-\cdots-y_{T-1}$, and $0^0 = 1$), since $m_{R+1}+\cdots+m_N = n_1$. With this notation we can define a second posterior $\rho'_{n,T} = \rho'_{n,T,x}$ given by

$$\begin{aligned}
\rho'_{n,T}(d\mu) &= \frac{\mu'(x)\rho_{0,T}(dx)}{\int_{Q_T}\mu'(x)\rho_{0,T}(dx)} \\
&= \frac{1}{Z'}\mathbf{1}_{Q_T} y_1^{m_1}\cdots y_R^{m_R}(1-y_1-\cdots-y_R)^{n_1}dy_1\cdots dy_{T-1},
\end{aligned} \tag{13}$$

with $Z'$ as the normalizing factor. Note that if each state has been observed at least twice (i.e, $R = T$) then the posteriors coincide; since this occurs with a probability which tends to one in $n$, we are really interested in small to moderate size samples.

For any given $T$, we start from computing the expected probability of each state under the posterior.

**Lemma 2.1.**

For any $\lambda > 0$ we have:

$$E_{\rho_{n,T,\lambda}}(y_i) = \frac{m_i + \lambda}{T\lambda + n}, \qquad i = 1, \ldots, N \tag{14}$$

$$E_{\rho_{n,T,\lambda}}(y_i) = \frac{\lambda}{T\lambda + n}, \qquad i = N+1, \ldots, T \tag{15}$$

$$E_{\rho'_{n,T}}(y_i) = \frac{m_i + 1}{T + n}, \qquad i = 1, \ldots, R \tag{16}$$

$$E_{\rho'_{n,T}}(y_i) = \frac{n_1}{(T-R)(T+n)} + \frac{1}{T+n}, \qquad i = R+1, \ldots, T \tag{17}$$

PROOF: The classical beta integral gives, for any pair $a, b \geq 0$ and any $x \in [0.1]$

$$\int_0^{1-x} y^a (1-x-y)^b dy = \frac{\Gamma(a+1)\Gamma(b+1)}{\Gamma(a+b+2)} (1-x)^{(a+b+1)}, \tag{18}$$

where $\Gamma$ is the gamma function.

To compute $E_{\rho_n,T,\lambda}(y_i)$ notice that the distribution of $\rho_{n,T,\lambda}$ remains unchanged under a permutation of the subscripts so that we can compute the integrals in any order we like: it turns out to be convenient to integrate with respect to $y_i$ last. Therefore, the next calculation performed for $i = 1$ is valid for all $i = 1, \ldots, T-1$.

We have

$$\begin{aligned} &E\rho_{n,T,\lambda}(y_1) \\ &\quad = \frac{1}{Z} \int_{Q_T} y_1^{m_1+\lambda} y_2^{m_2+\lambda-1} \ldots y_{T-1}^{m_{T-1}+\lambda-1} (1 - y_1 - \ldots - y_{T-1})^{m_T+\lambda-1} dy_1 \ldots dy_{T-1}. \end{aligned}$$

Let $\rho_i = m_i + \lambda - 1$ for $i = 1, \ldots, T$, and $\tilde{\rho}_i = \rho_i + \delta_{i,1}$, where $\delta$ is the Kronecker function; moreover, for all $i = 1, \ldots, T-1$, let

$$G_i = \frac{\Gamma(1+\rho_i)\Gamma(1+\sum_{s=i+1}^{T}\rho_s + T - 1 - i)}{\Gamma(2+\sum_{s=i}^{T}\rho_s + T - 1 - i))}.$$

Further, let $\tilde{G}_i$ be as $G_i$ with $\rho_s$ replaced by $\tilde{\rho}_s$ and let

$$\tilde{I}(i) = \int_{Q_i} y_1^{\tilde{\rho}_1} y_2^{\tilde{\rho}_2} \ldots y_i^{\sum_{s=i+1}^{T}\tilde{\rho}_i + T - s} dy_1 \ldots dy_i. \tag{19}$$

so that $E\rho_{n,T,\lambda}(y_1) = \tilde{I}(T-1)/I(T-1)$

Then

$$\begin{aligned}
\tilde{I}(T-1) &= \frac{\Gamma(1+\tilde{\rho}_{T-1})\Gamma(1+\tilde{\rho}_T)}{\Gamma(2+\tilde{\rho}_{T-1}+\tilde{\rho}_T)} I(T-2) \\
&= \tilde{G}_{T-1}\frac{\Gamma(1+\tilde{\rho}_{T-2})\Gamma(1+\tilde{\rho}_{T-1}+\tilde{\rho}_T+1)}{\Gamma(2+\tilde{\rho}_{T-2}+\tilde{\rho}_{T-1}+\tilde{\rho}_T+1)} I(T-3) \\
&= \prod_{s=1}^{T-1}\tilde{G}_s
\end{aligned}$$

with $I(T-1) = \prod_{s=1}^{T-1} G_s$.

Notice that $\tilde{G}_s = G_s$ for all $s$ except $s = 1$, so that

$$\begin{aligned}
E_{\rho_{n,T,\lambda}}(y_1) &= \frac{\tilde{I}(T-1)}{I(T-1)} \\
&= \prod_{s=1}^{T-1}\frac{\tilde{G}_s}{G_s} \\
&= \frac{\Gamma(\rho_1+2)}{\Gamma(\sum_{s=1}^{T}\rho_s + 1 + T)}\frac{\Gamma(\sum_{s=1}^{T}\rho_s + T)}{\Gamma(\rho_1+1)} \\
&= \frac{1+\rho_1}{\sum_{s=1}^{T}\rho_s + T} \\
&= \frac{m_1+\lambda}{\sum_{i=1}^{T} m_i + T\lambda} = \frac{m_1+\lambda}{T\lambda + n}
\end{aligned}$$

Recalling that $m_i = 0$ for $i \geq N$ we get the result for all $i < T$. Finally

$$\begin{aligned} E_{\rho_{n,T,\lambda}}(y_T) &= 1 - \sum_{i=1}^{T-1} E_{\rho_{n,T,\lambda}}(y_i) \\ &= 1 - \sum_{i=1}^{N} \frac{m_i + \lambda}{T\lambda + n} - (T - N - 1)\frac{\lambda}{T\lambda + n} \\ &= 1 - \frac{n + N\lambda + (T - N - 1)\lambda}{T\lambda + n} = \frac{\lambda}{T\lambda + n} \end{aligned}$$

which yields the result for the regular Bayesian.

In the modified Bayesian the calculation can be carried out in the same manner, with some modifications at the end. Although we are interested in the case $\lambda = 1$, we follow the same strategy as in the regular Bayesian, and obtain the result for all $\lambda > 0$. In this part of the proof we dente all quantities with a prime. For $i \leq R$, we have:

$$\begin{aligned} E_{\rho'_{n,T}}(y_i) &= E_{\rho'_{n,T}}(y_1) \\ &= \frac{1}{Z'} \int_{Q_T} y_1^{m_1+\lambda} y_2^{m_2+\lambda-1} \dots y_R^{m_R+\lambda-1} (1 - y_1 - \dots - y_R)^{n_1} y_{R+1}^{\lambda-1} \dots y_{T-1}^{\lambda-1} dy_1 \dots dy_{T-1} \\ &= \frac{1}{I'(T-1)} \tilde{I}'(T-1), \end{aligned}$$

with $I'(i)$ defined as in (**??**) with all $\rho$ replaced by $\rho'$, such that $\rho'_i = \rho_i$ and $\tilde{\rho}'_i = \tilde{\rho}_i$ for all $i = 1, \dots, R$ and $i = N+1, \dots, T-1$, but with $\tilde{\rho}'_i = \rho'_i = \lambda - 1$ for $i = R+1, \dots, N$.

Next, for $i = R+1, \dots, T-1$, let $G'_i$ be as $G_i$ with the needed primes, i.e.:

$$G'_i = \frac{\Gamma(1 + \rho'_i)\Gamma(1 + \sum_{s=i+1}^{T} \rho'_s + T - 1 - i)}{\Gamma(2 + \sum_{s=i}^{T} \rho'_s + T - 1 - i))};$$

and for $i = 1, \dots, R$, let instead

$$G'_i = \frac{\Gamma(1 + \rho'_i)\Gamma(1 + n_1 + \sum_{s=i+1}^{T} \rho'_s + T - 1 - i)}{\Gamma(2 + n_1 + \sum_{s=i}^{T} \rho'_s + T - 1 - i))}.$$

As before, let $\tilde{G}'_i$ be as $G'_i$ with $\rho'_s$ replaced by $\tilde{\rho}'_s$.

We have again that $\tilde{I}'(T-1) = \prod_{s=1}^{T-1} \tilde{G}'_s$ and that $\tilde{G}'_s = G'_s$ for all $s$ except $s = 1$, so that

$$\begin{aligned} E_{\rho'_{n,T,\lambda}}(y_1) &= \frac{\tilde{I}'(T-1)}{I'(T-1)} \\ &= \prod_{s=1}^{T-1} \frac{\tilde{G}'_s}{G'_s} \\ &= \frac{\Gamma(\rho'_1+2)}{\Gamma(n_1+\sum_{s=1}^{T}\rho'_s+1+T)} \frac{\Gamma(n_1+\sum_{s=1}^{T}\rho'_s+T)}{\Gamma(\rho'_1+1)} \\ &= \frac{1+\rho'_1}{n_1+\sum_{s=1}^{T}\rho_s+T} \\ &= \frac{m_1+\lambda}{n_1+\sum_{i=1}^{R} m_i+T\lambda} = \frac{m_1+\lambda}{T\lambda+n}. \end{aligned}$$

For $\lambda = 1$ we get (**??**).

For $i = R+1, \ldots, T-1$ the expected value becomes

$$\begin{aligned} E_{\rho'_{n,T,\lambda}}(y_i) &= E_{\rho'_{n,T,\lambda}}(y_{R+1}) \\ &= \frac{1}{Z'} \int_{Q_T} y_1^{m_1+\lambda-1} y_2^{m_2+\lambda-1} \ldots y_R^{m_R+\lambda-1} (1-y_1-\ldots-y_R)^{n_1} y_{R+1}^{\lambda} \ldots y_{T-1}^{\lambda-1} dy_1 \ldots dy_{T-1} \\ &= \frac{1}{K(T-1)} \tilde{K}(T-1). \end{aligned}$$

This time we let $\rho$ as before, but $\tilde{\rho}'_i = \tilde{\rho}_i - \delta_{i,1} + \delta_{i,R+1}$. Then

$$K(T-1) = \prod_{i=1}^{T-1} \bar{G}'_i$$

with $\bar{G}'_i = G'_i$ for all $i = R+2, \ldots, T-1$,

$$\bar{G}'_{R+1} = \frac{\Gamma(1+\rho_{R+1}+1)\Gamma(1+\sum_{s=R+2}^{T}\rho_s+T-R-2)}{\Gamma(2+\sum_{s=R+1}^{T}\rho_s+1+T-R-2)};$$

and, for $i = 1, \ldots R$,

$$\bar{G}'_i = \frac{\Gamma(1+\rho_i+1)\Gamma(1+n_1+\sum_{s=i+1}^{T}\rho_s+T-i-1)}{\Gamma(2+n_1+\sum_{s=i}^{T}\rho_s+1+T-i-1)}.$$

We no longer have $\bar{G}'_i = G'_i$, but two terms cancel in $\bar{G}'_i$ and $\bar{G}'_{i-1}$ for all $i = 2, \ldots, R$ so that, since $\Gamma(a+1) = a\Gamma(a)$,

$$\begin{aligned} E_{\rho'_{n,T,\lambda}}(y_i) &= \frac{K(T-1)}{Z'} \\ &= \frac{(\rho_{R+1}+1)(1+n_1+\sum_{s=R+1}^{T}\rho_s+(T-R)\lambda-1)}{(2+n_1+\sum_{s=1}^{T}\rho_s+T\lambda-2)(2+\sum_{s=R+1}^{T}\rho_s+(T-R)\lambda-2)} \\ &= \frac{n_1+(T-R)\lambda}{(n+T\lambda)(T-R)\lambda}, \end{aligned}$$

for $i = R+1, \ldots, T-1$.

To complete the result notice that

$$\begin{aligned} E_{\rho'_{n,T,\lambda}}(y_T) &= 1-\sum_{i=1}^{T-1} E_{\rho'_{n,T,\lambda}}(y_i) \\ &= 1-\sum_{i=1}^{R}\frac{m_i+1}{n+T}-(T-R-1)\frac{n_1+(T-R)\lambda}{(n+T\lambda)(T-R)\lambda} \\ &= \frac{n_1+(T-R)\lambda}{(n+T\lambda)(T-R)\lambda}. \end{aligned}$$

For $\lambda = 1$ this completes the proof.

□

## 2.1 Estimates of the number of species

We now move on to the estimation of $T$ by explicitly writing the expression for the expected unobserved probability $U$. In addition, according to the rationale that states observed once belong to the same class as those not observed, we also compute the expectation of the "unobserved" probability $U' = \sum_{i:m_i \leq 1} p_i$. Let $\rho'_{n,T} = \rho'_{n,T,1}$.

**Corollary 2.2** For $T \geq N$, we have

$$
\begin{aligned}
E_{\rho_{n,T,\lambda}}(U) &= \frac{(T-N)\lambda}{n+T\lambda}, \\
E_{\rho'_{n,T}}(U) &= \frac{T-N}{(n_1+T-N)}\frac{(2n_1+T-N)}{(T+n)}, \\
E_{\rho_{n,T,\lambda}}(U') &= \frac{n_1+(T-R)\lambda}{n+T\lambda}, \\
E_{\rho'_{n,T}}(U') &= \frac{(T-N+2n_1)}{n+T}
\end{aligned}
\tag{20}
$$

Proof

We have

$$
E_{\rho_{n,T,\lambda}}(U) = \sum_{i=N+1}^{T} \frac{\lambda}{n+T\lambda} = \frac{(T-N)\lambda}{n+T\lambda};
$$

$$
\begin{aligned}
E_{\rho'_{n,T}}(U) &= \sum_{i=N+1}^{T} \left(\frac{n_1}{(n_1+T-N)(T+n)} + \frac{1}{T+n}\right) \\
&= \frac{(2n_1+T-N)(T-N)}{(n_1+T-N)(T+n)};
\end{aligned}
$$

$$
E_{\rho_{n,T,\lambda}}(U') = \sum_{i=R+1}^{N} \frac{m_i+\lambda}{n+T\lambda} + \sum_{i=N+1}^{T} \frac{\lambda}{n+T\lambda} = \frac{n_1+(T-R)\lambda}{n+T\lambda};
$$

$$
\begin{aligned}
E_{\rho'_{n,T}}(U') &= \sum_{i=R+1}^{T} \left(\frac{n_1+T-R}{(T-R)(T+n)}\right) \\
&= \frac{(T-N+2n_1)}{n+T}
\end{aligned}
$$

□

In the next Lemma we collect some properties of the above expected values seen as functions of $T$, in order to determine which ones can be used to determine estimators of $T$.

**Lemma 2.3** For fixed $n$, $N$ and $n_1$ we have

(i) $E_{\rho_{n,N,\lambda}}(U) = E_{\rho'_{n,N}}(U) = 0$,
$E_{\rho_{n,N,\lambda}}(U') = \frac{n_1+(N-R)\lambda}{n+N\lambda}$,
$E_{\rho'_{n,N}}(U') = \frac{2n_1}{n+N}$;

(ii) $E_{\rho_{n,T,\lambda}}(U)$ is strictly increasing in $T$ for all $T \geq N$;
(iii) $E_{\rho'_{n,T}}(U)$ is strictly increasing in $T$ for all $T \geq N$;
(iv) if $n_1 < n$ then $E_{\rho_{n,T,\lambda}}(U')$ is strictly increasing in $T$ for all $T \geq N$;
(v) $E_{\rho'_{n,T}}(U')$ is strictly increasing in $T$ for all $T \geq N$
(vi) $\lim_{T\to\infty} E_{\rho_{n,T,\lambda}}(U) = \lim_{T\to\infty} E_{\rho'_{n,T}}(U)$
$= \lim_{T\to\infty} E_{\rho_{n,T,\lambda}}(U') = \lim_{T\to\infty} E_{\rho'_{n,T}}(U') = 1$

Proof (i) and (vi) follow immediately from the expressions in Lemma 2.2.

Then observe that for any $a, b \in \mathbb{R}$, $b > a$, and any $\lambda > 0$ the function $\frac{T\lambda+a}{T\lambda+b}$ is strictly increasing in T.

(ii) then follows for $b = n > -N\lambda = a$.

(iv) follow for $b = n > n_1 - R\lambda = a$, since $n > n_1$, except when $n_1 = n$, in which case, in fact, $R = 0$.

As to (iii), we can write $\frac{(2n_1+T-N)(T-N)}{(n_1+T-N)(T+n)} = f_1(T)f_2(T)$, with $f_1(T) = \frac{(T-N)}{(n_1+T-N)}$. We then apply the same reasoning as above with $\lambda = 1$. Thus, $f_1'(T) > 0$, since we can take $b = n_1 - N > -N = a$, except when $n_1 = 0$ in which case $f_1(T) = 1$. Also, $f_2'(T) > 0$ by observing that $b = n > 2n_1 - N = a$, except when $n = N = n_1$ in which case $f_2(T) = 1$. Therefore, $(f_1 f_2)' > 0$ except when $n = N = n_1 = 0$, which is impossible.

Finally, (v) follows easily by similar arguments.

□

From the properties listed in Lemma 2.3, the expected values involving $U'$ are not suitable for determing $T$ when equated to the Turing-Good estimator, because of their behaviour around $T = N$. For the expected values involving $U$, instead, the properties above guarantee that there is a unique solution of equations of the form $E_{\rho_{n,T}}(U) = \hat{U}$ and that this is some $\hat{T} \in [N, +\infty]$. Such roots, or rather some integer approximation, will be taken as our estimators.

Note that, by (i)-(iii), the root $\hat{T}$ of any such equation satisfies $\hat{T} = N$ iff $\hat{U} = 0$ and that, by (ii), (iii) and (vi) $\hat{T} = +\infty$ iff $\hat{U} = 1$.

**Theorem 2.4**.

Let $\hat{U}$ be a real number in $[0, 1]$. Then the unique solution in $[N, +\infty]$ of the equation $E_{\rho_{n,T,\lambda}}(U) = \hat{U}$ is

$$\frac{N}{1-\hat{U}} + \frac{n\hat{U}}{(1-\hat{U})}\frac{1}{\lambda}. \tag{21}$$

and the unique solution of $E_{\rho'_{n,T}}(U) = \hat{U}$ in $[N, \infty]$ has the form

$$\begin{aligned} &\frac{(N-n_1)(2-\hat{U}) + n\hat{U} + \sqrt{4n_1^2(1-\hat{U}) + (\hat{U})^2(n+N-n_1)^2}}{2(1-\hat{U})} \\ &= \frac{N}{(1-\hat{U})} + \frac{n\hat{U}}{(1-\hat{U})}\gamma^2 \end{aligned} \tag{22}$$

Furthermore, $1 > \gamma^2 \geq 0$ and $\gamma^2 = 0$ iff all states have been observed exactly once (which is to say, $n_1 = N = n$).

Finally, $\hat{T} = N$ and $\hat{T}' = N$ iff $\hat{U} = 0$, and $\hat{T} = +\infty$ and $\hat{T}' = +\infty$ iff $\hat{U} = 1$.

Proof.

(**??**) is trivial and we let $\hat{T}(\lambda)$ to equal the integer part of $\frac{N}{(1-\hat{U})} + \frac{n\hat{U}}{(1-\hat{U})}\frac{1}{\lambda}$

The equation $E_{\rho'_{n,T}}(U) = \hat{U}$ has solutions

$$
\begin{aligned}
& T'_{+,-} = \\
& \left( \frac{(N-n_1)(2-\hat{U}) + n\hat{U} \pm \sqrt{4(n_1)^2(1-\hat{U}) + (\hat{U})^2(n+N-n_1)^2}}{2(1-\hat{U})} \right) \\
= \; & N + \left( \frac{n_1(-2+\hat{U}) + (n+N)\hat{U} \pm \sqrt{4(n_1)^2(1-\hat{U}) + (\hat{U})^2(n+N-n_1)^2}}{2(1-\hat{U})} \right).
\end{aligned} \tag{23}
$$

Note that

$$
\begin{aligned}
& (4(n_1)^2(1-\hat{U}) + (\hat{U})^2(n+N-n_1)^2) - (n_1(-2+\hat{U}) + (n+N)\hat{U})^2 \\
& \qquad = 4n_1(n+N)\hat{U}(1-\hat{U}) \quad \geq 0,
\end{aligned} \tag{24}
$$

with strict inequality for non trivial $U$, which is to say different from 0 or 1. This implies that $T'_- \leq N$ and $T'_+ \geq N$, with strict inequalities for nontrivial $T$, so that $T'_+$ is the unique solution of $E_{\rho'_{n,N}}(U) = \hat{U}$ in $[N, +\infty)$ and we then let $\hat{T}'$ equal the integer part of $T'_+$. By Lemma 2.3 such a solution equals $N$ if and only if $\hat{U} = 0$ and can be taken to equal $+\infty$ if and only if $\hat{U} = 1$.

We can now write

$$
T'_+ = \frac{N}{(1-\hat{U})} + \frac{n\hat{U}}{(1-\hat{U})}\gamma^2 \tag{25}
$$

with

$$
\gamma^2 = \frac{(n-N+n_1)\hat{U} - 2n_1 + \sqrt{4(n_1)^2(1-\hat{U}) + (\hat{U})^2(n+N-n_1)^2}}{2n\hat{U}}.
$$

Simple calculations show that $\gamma^2 = (n-n_1)/n$ for $\hat{U} = 1$ and $\lim_{\hat{U}\to 0}\gamma^2 = (n-N)/2n$.

Note also that if all states have been observed exactly once in the sample this means that $n_1 = N = n$, and in such case a simple calculation shows that $\gamma^2 = 0$. Moreover, $\gamma^2 \geq 0$, as

follows from

$$(4(n_1)^2(1-\hat{U}) + (\hat{U})^2(n+N-n_1)^2 - ((n-N+n_1)\hat{U} - 2n_1)^2 \\ = 4\hat{U}(n_1(n-N) + n\hat{U}(N-n_1)) \geq 0,$$

since $n \geq N$ and $N \geq n_1$. Furthermore, the above inequality is strict unless one of two things happen: (1) $\hat{U} = 0$, and in such case $\gamma^2 > 0$ unless $n = N$ which already implies that all states have been observed once, or (2) when $n = N$ and $N = n_1$, implying again that all states are observed once.

Finally, $\gamma^2 < 1$ since

$$1 - \gamma^2 = \frac{(n+N-n_1)\hat{U} + 2n_1 - \sqrt{4(n_1)^2(1-\hat{U}) + (\hat{U})^2(n+N-n_1)^2}}{2n\hat{U}}$$

and

$$((n+N-n_1)\hat{U} + 2n_1)^2 - (4(n_1)^2(1-\hat{U}) + (\hat{U})^2(n+N-n_1)^2) \\ = 4\hat{U}n_1(n+N) > 0;$$

in fact, the last inequality is strict unless $\hat{U} = 0$, but in that case we have already seen that the limit of $\gamma^2$ is strictly less than 1.

□

The integer part of $\hat{T}(1)$ and $\hat{T}'$ of the two roots in Theorem 2.4 are taken as our estimates of $T$ and will be called regular Bayesian and modified Bayesian estimator, respectively. These expressions coincide with the Chao-Lee estimator for the values $\gamma^2 = 1$ and for the given $\gamma^2$, respectively.

In Section 5 we test the two estimators $\hat{T} = \hat{T}(1)$ and $\hat{T}'$ with smoothed Turing-Good $\hat{U}$ on some explicit data. As expected, $\hat{T}$, which also corresponds to $\hat{T}'$ with $\gamma = 1$, is too simple and does not perform well in most samples; on the other hand, $\hat{T}'$ turns out to be the best available estimator, performing particularly well for far from uniform species distributions.

## 2.2 Estimates of species distribution

We now turn to the estimation of the species distribution. A reasonable estimate for this is the posterior average probability of each species as computed in Lemma 2.1 under either $\rho_n$ or $\rho'_n$ taking the corresponding estimated value $\hat{T}$ and $\hat{T}'$, respectively, as a value for $T$.

For the first estimator, taking $\hat{T}$ as in (**??**), this leads to

$$E_{\rho_n,\hat{T},\lambda}(y_i) = \frac{m_i+\lambda}{\hat{T}\lambda+n} = \frac{(m_i+\lambda)(1-\hat{U})}{n+N\lambda}, \qquad i=1,\dots,N, \tag{26}$$

and

$$E_{\rho_n,\hat{T},\lambda}(y_i) = \frac{\lambda}{\hat{T}\lambda+n} = \frac{\lambda(1-\hat{U})}{n+N\lambda}, \qquad i=N+1,\dots,\hat{T}_1 \tag{27}$$

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

Note that the values in (**??**) are close to the unbiased estimator $m_i/n$ of the probability of the $i$-th species. The above estimation also constitutes a mixture of the Laplace add-$\lambda$ and Turing-Good estimators: it is in fact 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 having assigned probability $\hat{U}$ to the event that we will observe a new species; the estimate of each of the $N$ species is then reduced by the factor $1-\hat{U}$ to compensate for this. This is likely to be a sensible way to make the attenuation of the Laplace estimator (see Orlitsky et al (2003)) finite. If no better estimation of $\lambda$ is available, one can always use $\lambda = 1$.

For the second estimator, taking $\hat{T}'$ as in (**??**), this gives for $i = 1,\dots,R$

$$\begin{aligned} E_{\rho'_n,\hat{T}'}(y_i) &= \frac{m_i+1}{\hat{T}'+n} \\ &= \frac{2(1+m_i)}{(2-\hat{U})(n+N-n_1)+\sqrt{4(n_1)^2(1-\hat{U})+(\hat{U})^2(n+N-n_1)^2}} \end{aligned}$$

and for $i = R+1,\dots,\hat{T}'$, with $a = 4(n_1)^2(1-\hat{U})+(\hat{U})^2(n+N-n_1)^2$,

$$\begin{aligned} E_{\rho'_n,\hat{T}'}(y_i) &= \frac{n_1}{(n_1+\hat{T}'-N)(\hat{T}'+n)} + \frac{1}{(\hat{T}'+n)} \\ &= \frac{2(1-\hat{U})(\sqrt{a}+2n_1-2n_1\hat{U}+R\hat{U})}{(\sqrt{a}+2n+2R-2n\hat{U}-R\hat{U})(\sqrt{a}+R\hat{U})}. \end{aligned}$$

Note that this value is close to a natural estimation in our scheme: we estimated that there are $n_1+\hat{T}'-N$ species which have been observed less than twice; these together share

a probability that we estimate of the order of $n_1/n$, and since there is no further element to distinguish one species from the other, a natural estimate for their probability is $\frac{1}{n_1+\hat{T}'-N}\frac{n_1}{n}$, close to the expression we get.

## 3. Estimates of the distribution of the unobserved mass $U$

In this section we give a Bayesian interpretation of $P(U > \epsilon)$ and estimate it by using the value of $T$ previously estimated by our $\hat{T}$ or $\hat{T}'$. The idea is to assume that $T$ is known and then define $P(U > \epsilon)$ as the probability under the posterior of the distributions which give weight greater than or equal to $\epsilon$ to the unobserved species. When we have such an expression we then replace $T$ by one of its estimates: the resulting function is our estimate of the distribution of $U$. Note that we assume continuity in all the parameters involved so that we do not distinguish between $P(U > \epsilon)$ and $P(U \geq \epsilon)$.

There are, however, various possible choices for the quantities we intend to use: the posterior can be generated in the standard Bayesian form, or with our modified version; as set of unobserved states one can take those not observed at all, or those observed zero times or once, and, finally, we have two estimators for $T$. However, if we want the estimated distribution function to be roughly centered in $\hat{U}$ we should use $U = \{i : L(i) = 0\}$, and, for each likelihood, the estimation of $T$ derived with that likelihood. These restrictions give rise to only two forms of the estimated distribution of $U$: one is

$$
\begin{aligned}
(28) \qquad P_\lambda(U > \epsilon) &= \rho_{n,\hat{T},\lambda}(U > \epsilon) \\
&= \rho_{n,\hat{T},\lambda}\left(\mu = (p_1, \ldots, p_{\hat{T}}) : \mu(U) = \sum_{i:L_n(i)=0} p_i > \epsilon\right),
\end{aligned}
$$

of which we only consider the value $P(U > \epsilon) = P_1(U > \epsilon)$ for $\lambda = 1$; and the other is the modified version

$$
\begin{aligned}
(29) \qquad P'(U > \epsilon) &= \rho'_{n,\hat{T}'}(U > \epsilon) \\
&= \rho'_{n,\hat{T}'}\left(\mu = (p_1, \ldots, p_{\hat{T}'}) : \mu(U) = \sum_{i:L_n(i)=0} p_i > \epsilon\right).
\end{aligned}
$$

Since $\hat{T}'$ is a very good estimator, $P'$ is likely to produce an effective estimation of the distribution of $U$ while (**??**) with $\lambda = 1$ is not likely to provide a good estimation. However, there are several reasons to include first the explicit form of (**??**): it gives rough preliminary estimates; it is easier to implement; the explicit formula for (**??**) can be written in terms of that for (**??**); but, above all, the expression we get for (**??**) is computationally too heavy and already for the mid size example from the Carothers data we are using in section 5 we have to resort to a mixed method in which really only (**??**) for $\lambda = 1$ is computed.

**Lemma 3.1**.

For $\hat{T}$ integer

$$\Psi_{\hat{T}}(\epsilon) = P(U > \epsilon) = (1-\epsilon)^{\hat{T}+n-1} \sum_{i=1}^{\hat{T}-N} \left(\frac{\epsilon}{1-\epsilon}\right)^{\hat{T}-N-i} \frac{\Gamma(\hat{T}+n)}{(\hat{T}-N-i)!\Gamma(i+n+N)} \tag{30}$$

and

$$\begin{aligned} \Psi'_{\hat{T}'} = P'(U > \epsilon) = & \sum_{i=1}^{(\hat{T}'-N)} \frac{\Gamma(\hat{T}'-N)\Gamma(\hat{T}'-N+n_1)}{\Gamma(\hat{T}'-N-i+1)\Gamma(i+n_1)} \sum_{j=0}^{(i+n_1-1)} \binom{i+n_1-1}{j} (-1)^j \\ & \times \left( \sum_{s=1}^{i+2n_1-j} \epsilon^{\hat{T}'-N+2n_1-s} (1-\epsilon)^{R+n-n_1+s-1} \frac{\Gamma(i+2n_1-j)}{\Gamma(i+2n_1-j-s+1)\Gamma(i+n-n_1+R)} \right) \\ & \times \frac{\Gamma(\hat{T}'-N+n-n_1+R)}{\Gamma(\hat{T}'-N+2n_1)}. \end{aligned} \tag{31}$$

Proof.

We start thus from (**??**), which is to say (**??**) with $\lambda = 1$; if $\hat{T} > N$ then by the expression of the standard Bayesian posterior we get

$$\begin{aligned} P(U \geq \epsilon) &= \rho_{n,\hat{T},1}(U > \epsilon) \\ &= \frac{1}{Z} \int\limits_{\substack{y_i \geq 0 \\ \sum_{i=1}^{N} y_i \leq 1-\epsilon, \quad \sum_{i=1}^{\hat{T}-1} y_i \leq 1}} y_1^{m_1} \cdots y_N^{m_N} dy_1 \cdots dy_N \cdots dy_{\hat{T}-1}, \end{aligned} \tag{32}$$

where $Z$ is the usual normalizing factor, this time with $\lambda = 1$ and $T$ replaced by $\hat{T}$.

If $\hat{T} = N$ then we conclude that all states have been observed and that, therefore, $P(U > \epsilon) = 0$ for all $\epsilon$.

It is possible to give an explicit expression for the r.h.s. as follows.

Let $I_\epsilon$ denote the integral in the r.h.s. of the above equation, so that $P(U \geq \epsilon) = I_\epsilon / Z$ and let $K = \hat{T} - N$. We now can get an explicit expression for $I_\epsilon$ by first integrating with respect to the variables $y_{N+1}, \ldots, y_{T-1}$; this gives

$$I_\epsilon = \frac{1}{(K-1)!} \int\limits_{\substack{y_i \geq 0 \\ \sum_{i=1}^{N} y_i \leq 1-\epsilon}} y_1^{m_1} \ldots y_N^{m_N} (1 - y_1 - \cdots - y_N)^{K-1} dy_1 \ldots dy_N. \tag{33}$$

To get a simple expression we now want to reduce to gamma integrals and this can be achieved by integrating by parts several times until the expression to the power $(K-1)$ disappears. We then arrive at

$$\begin{aligned} I_\epsilon &= \frac{1}{(K-1)!} \int\limits_{\substack{y_i \geq 0 \\ \sum_{i=1}^{N-1} y_i \leq 1-\epsilon}} y_1^{m_1} \ldots y_{N-1}^{m_{N-1}} \cdot \\ &\left( \sum_{i=1}^{K} \epsilon^{K-i} \frac{(K-1)! m_N!}{(K-i)!(m_N+i)!} (1 - \epsilon - y_1 - \ldots y_{N-1})^{m_N+i} \right) dy_1 \ldots dy_{N-1}. \end{aligned} \tag{34}$$

Now, for each term in the sum we can perform $N-1$ gamma integrals and simplify:

$$\begin{aligned}
&\int\limits_{\substack{y_i\geq 0\\ \sum_{i=1}^{N-1} y_i\leq 1-\epsilon}} y_1^{m_1}\dots y_{N-1}^{m_{N-1}}(1-\epsilon-y_1-\dots y_{N-1})^{m_N+i}dy_1\dots dy_{N-1}\\
&\quad=(1-\epsilon)^{i+\sum_{j=1}^N m_j+N-1}\prod_{r=1}^{N-1}\frac{\Gamma(1+m_{N-r})\Gamma(1+i+\sum_{j=N-r+1}^N m_j+r-1)}{\Gamma(2+i+\sum_{j=N-r}^N m_j+r-1)}\\
&\quad=(1-\epsilon)^{i+\sum_{j=1}^N m_j+N-1}\frac{\Gamma(1+i+m_N)\prod_{j=1}^{N-1}\Gamma(1+m_j)}{\Gamma(i+\sum_{j=1}^N m_j+N)}\\
&\quad=(1-\epsilon)^{i+n+N-1}\frac{\Gamma(1+i+m_N)\prod_{j=1}^{N-1}\Gamma(1+m_j)}{\Gamma(i+n+N)}
\end{aligned}$$

so that

$$\begin{aligned}
(K-1)!I_\epsilon &= \sum_{i=1}^K \epsilon^{K-i}\frac{(K-1)!m_N!}{(K-i)!(m_N+i)!}(1-\epsilon)^{i+n+N-1}\frac{\Gamma(1+i+m_N)\prod_{j=1}^{N-1}\Gamma(1+m_j)}{\Gamma(i+n+N)}\\
&= (1-\epsilon)^{T+n-1}\sum_{i=1}^K(\frac{\epsilon}{1-\epsilon})^{K-i}\frac{(K-1)!m_N!}{(K-i)!\Gamma(i+n+N)}\prod_{j=1}^{N-1}\Gamma(1+m_j). \qquad (35)
\end{aligned}$$

Recall that $Z=I_0$ which, in the expression above, means $i=K$, so that, by the definition of $K$, we have

$$\frac{I_\epsilon}{Z} = (1-\epsilon)^{\hat{T}+n-1}\sum_{i=1}^{\hat{T}-N}(\frac{\epsilon}{1-\epsilon})^{\hat{T}-N-i}\frac{\Gamma(\hat{T}+n)}{(\hat{T}-N-i)!\Gamma(i+n+N)}$$

Notice that in (**??**), for any positive integers $s$ and $t$, with $N, K, n$ and $T$ replaced respectively by $s, t+1, \sum_{j=1}^s m_j$ and $s+t+1$, we have

$$\begin{aligned}
&\int\limits_{\substack{y_i\geq 0\\ \sum_{i=1}^{s} y_i\leq 1-\epsilon}} y_1^{m_1}\cdots y_s^{m_s}(1-y_1-\dots-y_s)^t dy_1\dots dy_s = (1-\epsilon)^{t+s+\sum_{j=1}^s m_j}\times\\
&\qquad\sum_{i=1}^{t+1}(\frac{\epsilon}{1-\epsilon})^{t+1-i}\frac{t!m_s!}{(t-i+1)!\Gamma(i+\sum_{j=1}^s m_j+s)}\prod_{j=1}^s\Gamma(1+m_j) \qquad (36)
\end{aligned}$$

In (**??**), we need to integrate over the region $\Sigma = \{y = (y_1, \ldots, y_{T-1}) \in Q_T : \sum_{i=1}^{N} y_i \leq 1 - \epsilon, \sum_{i=1}^{T-1} y_i \leq 1\}$ obtaining

$$
\begin{aligned}
P'(U > \epsilon) &= \rho'_{n,T'} \left( \mu = (p_1, \ldots, p_{\hat{T}'}) : \mu(U) = \sum_{i:L_n(i)=0} p_i \geq \epsilon \right) \\
&= \frac{1}{Z'} \int_{\Sigma} y_1^{m_1} \cdots y_R^{m_R} (1 - y_1 - \cdots - y_R)^{n_1} dy_1 \cdots dy_N \cdots dy_{\hat{T}'-1}, \qquad (37) \\
&= \frac{I'_\epsilon}{Z'}
\end{aligned}
$$

where $Z' = I'_0$.

This case is more involved. To compute $I'_\epsilon$ we carry out the first $K - 1 = T - 1 - N$ integrations, as before, with respect to the variables $y_{N+1}, \ldots, y_{T-1}$; this gives

$$
\begin{aligned}
(K-1)! \quad I'_\epsilon \\
&= \int_{\substack{y_i \geq 0 \\ \sum_{i=1}^{N} y_i \leq 1-\epsilon}} y_1^{m_1} \ldots y_R^{m_R} (1 - y_1 - \cdots - y_R)^{n_1} (1 - y_1 - \cdots - y_N)^{K-1} dy_1 \ldots dy_N. \qquad (38)
\end{aligned}
$$

Now we again integrate $K - 1$ times by parts with respect to $y_N$, then with respect to the $n_1 - 1$ variables $y_{N-1}, \ldots y_{R+1}$, expand in powers of $\epsilon$, and, finally, compute the last integrals

using (**??**). This leads to

$$
\begin{aligned}
(K-1)!\quad I'_\epsilon &= \int\limits_{\substack{y_i\geq 0\\ \sum_{i=1}^{N-1} y_i\leq 1-\epsilon}} y_1^{m_1}\dots y_R^{m_R}(1-y_1-\dots-y_R)^{n_1}\cdot \\
&\qquad \left(\sum_{i=1}^{K}\epsilon^{K-i}\frac{(K-1)!}{(K-i)!(i)!}(1-\epsilon-y_1-\ldots-y_{N-1})^i\right)dy_1\dots dy_{N-1} \\
&= \int\limits_{\substack{y_i\geq 0\\ \sum_{i=1}^{R} y_i\leq 1-\epsilon}} y_1^{m_1}\dots y_R^{m_R}(1-y_1-\dots-y_R)^{n_1}\cdot \\
&\qquad \left(\sum_{i=1}^{K}\epsilon^{K-i}\frac{(K-1)!}{(K-i)!(i)!}\frac{(1-\epsilon-y_1-\dots-y_R)^{i+n_1-1}}{(i+1)\dots(i+n_1-1)}\right)dy_1\dots dy_R \qquad (39)\\
&= \sum_{i=1}^{K}\epsilon^{K-i}\frac{(K-1)!}{(K-i)!(i+n_1-1)!}\sum_{j=0}^{i+n_1-1}\binom{i+n_1-1}{j}(-1)^j\epsilon^j \\
&\qquad \int\limits_{\substack{y_i\geq 0\\ \sum_{i=1}^{R} y_i\leq 1-\epsilon}} y_1^{m_1}\dots y_R^{m_R}(1-y_1-\dots-y_R)^{i+2n_1-1-j}dy_1\dots dy_R \\
&= \sum_{i=1}^{K}\frac{(K-1)!}{(K-i)!(i+n_1-1)!}\sum_{j=0}^{i+n_1-1}\binom{i+n_1-1}{j}(-1)^j \\
&\qquad \times \\
&\qquad\quad \left(\sum_{s=1}^{i+2n_1-j}\frac{\epsilon^{K+2n_1-s}}{(1-\epsilon)^{1-R-n+n_1-s}}\frac{(i+2n_1-1-j)!\quad m_R!}{(i+2n_1-j-s)!\Gamma(i+n-n_1+R)}\right)\times \\
&\qquad\qquad \prod_{j'=1}^{R}\Gamma(1+m_{j'})
\end{aligned}
$$

Taking $\epsilon = 0$ the only nonvanishing term is for $s = 2n_1 + K$; in this case, since $0 \leq j$ and $i \leq K$, so that $1 \leq s \leq i - j + 2n_1 \leq K - j + 2n_1 \leq K + n_1 = s$. Hence $i - j = K$, which implies that $i = K$ and $j = 0$. Therefore,

$$
I_0 = \frac{1}{(K+n_1-1)!}\frac{(K+2n_1-1)!\quad m_R!}{\Gamma(K+n-n_1+R)}\prod_{j'=1}^{R}\Gamma(1+m_{j'});
$$

then,

$$\begin{aligned}\frac{I'_\epsilon}{I'_0} &= \sum_{i=1}^{K} \frac{(K-1)!(K+n_1-1)!}{(K-i)!(i+n_1-1)!} \sum_{j=0}^{i+n_1-1} \binom{i+n_1-1}{j}(-1)^j \\ &\times \\ &\left( \sum_{s=1}^{i+2n_1-j} \frac{\epsilon^{K+2n_1-s}}{(1-\epsilon)^{1-R-n+n_1-s}} \frac{(i+2n_1-1-j)!}{(i+2n_1-j-s)!\Gamma(i+n-n_1+R)} \right) \times \\ &\frac{\Gamma(K+n-n_1+R)}{(K+2n_1-1)!} \end{aligned} \tag{40}$$

Substituting $K = \hat{T}' - N$ in $I'_\epsilon/I'_0$ we obtain the estimate $P'(U > \epsilon)$.

□

We now discuss the relation of our estimate of the distribution of $U$ with the one developed in Almudevar et al (2000), which is based on large deviations and a bootstrap method.

In Almudevar et al (2000) it is proven, by large deviations methods, that, for a suitable function $s^*(\epsilon)$, in a sample of size $n$,

$$P(U \geq \epsilon) \approx (1 - s^*(\epsilon))^n,$$

where $s^*(\epsilon)$ can be further estimated by

$$s_n^*(\epsilon) = \inf\{ \sum_{i:L_n(i)>0} L_n(i) : \sum_{i:L_n(i)>0} L_n(i) > \epsilon\}.$$

The first difference is a matter of interpretation: in Almudevar et al (2000), $P(U \geq \epsilon)$ represents the probability that in repeated samples from a fixed populations the total probability of the unobserved states exceeds $\epsilon$. This probability can clearly be computed if the population is known, or else estimated as above from the sample. Our approach, on the other hand, resembles more the probability that in testing different randomly selected populations the total probability of the unobserved states exceeds $\epsilon$. Clearly, in taking just one sample of one population both approaches can be considered.

The main difference between our method and the method in Almudevar et al (2000), however, is in the range of application of the two methods. Since the ABS method uses the sample to estimate the population distribution, it is not suitable with a small coverage (such as the 20% of the Carothers data): it makes more sense to use it when we think that most species have been observed and we want to estimate the low probability of still having missed some of them. In fact, $(1-s^*(\epsilon))^n \leq (1-\epsilon)^n$ is a very small number except when $\epsilon$ is very small. On the other hand, the method we develop here deals exactly with the opposite case, in which the coverage is largely incomplete. Thus, the two methods cover distinct possibilities, and we can use as discriminant the value of $n_1$: if $n_1 > 0$, coverage is likely to be incomplete and our method applies; otherwise, if $n_1 = 0$ coverage is more likely to be complete or almost so, and the ABS method applies (in this case, in fact, our estimate of $T$ is $N$, if we do not smooth $n_1$ and the estimated distibution of $U$ is just a trivial delta at 0).

## 4. Confidence intervals for $U$ and $T$

We now perform a second bootstrap in order to generate confidence intervals for $U$ and $T$. Note first that, by the methods of Section 3, we get an estimate of the distribution of $U$, so that given $1-\alpha > 0$ we can determine $\hat{U}_1$ and $\hat{U}_2$ such that the (estimated) probability that $U$ is in $[\hat{U}_1, \hat{U}_2]$ is greater than or equal to $1-\alpha$, i.e. $P_{\hat{T}}(U \in [\hat{U}_1, \hat{U}_2]) \geq \alpha$. The interval $[\hat{U}_1, \hat{U}_2]$ contains $\hat{U}_{TG}$, and one possible choice is to take it symmetric around $\hat{U}_{TG}$; when the modified Bayesian method is used to produce the interval we denote it by $[\hat{U}_1', \hat{U}_2']$. Any such interval can be considered as a confidence interval for $U$.

Confidence intervals for $U$ can in principle be obtained also from estimates of the error in the Turing-Good estimator $\hat{U}_{TG}$. However, the available estimates do not seem to provide useful intervals; for instance, the bounds provided in McAllester and Schapire (2000) are interesting only asymptotically in the sample size, and even for moderate size samples such as those in the Carothers data the bounds fall even outside the interval $[0, 1]$.

Next, we want to discuss confidence intervals for $T$. Recall that by Lemma 2.3, $E_{\rho'_{n,T}}(U)$ and $E_{\rho_{n,T}}(U)$ are strictly increasing in $T$. Then

$$[\hat{T}(\hat{U}_1), \hat{T}(\hat{U}_2)] = \{T : E_{\rho_{n,T}}(U) \in [\hat{U}_1, \hat{U}_2]\} \tag{41}$$

and

$$[\hat{T}'(\hat{U}_1'), \hat{T}'(\hat{U}_2')] = \{T : E_{\rho'_{n,T}}(U) \in [\hat{U}_1', \hat{U}_2']\} \tag{42}$$

where $\hat{T}(\hat{U})$ ($\hat{T}'(\hat{U})$, resp.) is the solution of $E_{\rho_{n,T}}(U) = \hat{U}$ ($E_{\rho'_{n,T}}(U) = \hat{U}$, respectively), can be considered confidence intervals for $T$ at level $\alpha$.

One possible choice for the interval $[\hat{U}_1, \hat{U}_2]$ is to take it symmetric around the Turing-Good estimate for the unobserved probability, so that the intervals $[\hat{T}(\hat{U}_1), \hat{T}(\hat{U}_2)]$ and $[\hat{T}'(\hat{U}_1), \hat{T}'(\hat{U}_2)]$ will certainly contain our pointwise estimates $\hat{T}$ and $\hat{T}'$, respectively. The functions that we use are taken from Theorem 2.4 we have

$$\hat{T}(\hat{U}) = \frac{N}{1-\hat{U}} + \frac{n\hat{U}}{1-\hat{U}}. \tag{43}$$

and

$$\hat{T}'(\hat{U}') = \frac{(N-n_1)(2-\hat{U}') + n\hat{U}' + \sqrt{4(n_1)^2(1-\hat{U}') + (\hat{U}')^2(n+N-n_1)^2}}{2(1-\hat{U}')} \tag{44}$$

Due to the numerical difficulties in computing the approximate distribution of $U$ using the modified Bayesian method, and thus in determining $\hat{U}'_i$, $i = 1, 2$, we actually develop a mixed method which turns out to be much more easily implementable, quite effective, and with extra flexibility. We already discussed how the function $\Psi'_{\hat{T}'}(\epsilon)$ (see Lemma 3.1) cannot be numerically computed even in mid size samples and how simple examples have indicated that it would give rise to a larger confidence interval than that obtained from $\Psi_{\hat{T}}(\epsilon)$. Thus $[\hat{U}'_1, \hat{U}'_2]$ contains $[\hat{U}_1, \hat{U}_2]$, or, for symmetric intervals around $\hat{U}_{TG}$, $\hat{U}'_2 = \hat{U}_{TG} + d'$ and $\hat{U}_2 = \hat{U}_{TG} + d$, with $d' > d$. In some examples $d'$ turns out to be a constant $\bar{c}$ times $d$, and $\Psi_{\hat{T}}(\epsilon)$ turns out to be an invertible function of $\epsilon$; we thus select some constant $\bar{c}$, which we take to be $\bar{c} = 2$ for convenience, and then use the following scheme:

- fix a confidence level $\alpha$,

- determine $d$ such that $\Psi_{\hat{T}}(\hat{U}_{TG} + d) - \Psi_{\hat{T}}(\hat{U}_{TG} + d) = 1 - \alpha$,

- let the confidence interval be $[\hat{T}'(\hat{U}_{TG} - \bar{c}d), \hat{T}'(\hat{U}_{TG} + \bar{c}d)]$

The extra flexibility of this method comes from the fact that in principle it can be applied even without any theoretical justification on the constant $\bar{c}$, just selecting a value which turns out to be effective in experimental examples.

The perfomance of the above confidence intervals for the Carothers data is evaluatedd in the next Section.

## 5. Data analysis

The estimators and confidence intervals discussed above for the number of unobserved species are computed here for three sets of data: the words used by Shakespeare, included mostly for curiosity, the Carothers data for Edinburgh taxi cabs and data from live trapping of cottontail rabbits. We do not try to estimate the unobserved probability, or the coverage, since its true value is not known, making it impossible to use the data as a test of the estimators. To make implementation simple, there is a software created precisely to calculate most of the estimators presented here, together with estimates on the error (see Colwell (2006)) but the calculations below are based on the formulas given here or in the original works (see Gandolfi and Sastri (2004), Section 4, for a review).

For the number of words used by and known to Shakespeare data containing the number of words used up to ten times can be found in Efron and Thisted (1976). Based on those numbers it is possible to compute several estimators, including those presented here. The number of words used by Shakespeare is 31,534 and the estimators are in the first column of Table 1. So, $\hat{T}_{MLE} = 31,534$, $\hat{T}_{TG} = 32,034$ ecc. Efron and Thisted (1976) gave their own estimate of 66,534, which looks a bit high compared to most estimators available.

The data in this example, as well as almost all others taken from real experiments, have the drawback that we do not really know the number we are trying to estimate, so this hardly constitutes a test for the estimators.

A very useful set of data is instead in Carothers (1973) for the taxi cabs in Edinburgh. The data consists of records of taxi cabs observed at different times and locations, and while they were meant to study estimation methods for the population size in capture-recapture experiments, they have been adapted to our present situation by interpreting every taxi cab as a different species and each observation as a different member of that species in the sample. An observation does not alter the probability of future observations, so this data could costitutes a very explicit example of the results of drawing with replacement from a population with different species. One of the advantages of the data is that it is entirely published and thus calculations based on it are fully reproducible.

The data is divided into several experiments, denoted as A $\alpha$, A$\beta$ etc. to B$\gamma$. From the way is has been sampled, the data somehow goes from more uniform to a less uniform distribution, with a drastic change in sampling technique between the A and the B data. So, really the estimators to be used for uniform distributions are expected to perform poorly in the later samples. Tables 1 and 2 report the performance of several estimators on the data for the various experiments. The estimators considered are: the maximum likelihood

$T_{MLE}$ (see Lewontin and Prout (1956)), the Turing-Good $T_{TG}$, the $i$-th Jack-Knife estimator $T_{JKi}$ (see Burnahm and Overton (1979)) for $i = 1, \ldots, 6$, the first Chao estimator $T_{C1}$ (see Chao (1984)), the Bias-Corrected Chao estimator $T_{C2}$ (see Colwell (2006)), the first and second Chao-Lee estimators $T_{CL1}$ and $T_{CL2}$ (see Chao and Lee (1992)), the abundance based coverage estimator $T_{ACE}$ (see Colwell (2006)), which for the Carothers data coincides with $T_{CL1}$, and our first and second Bayesian estimators $\hat{T}$ and $\hat{T}'$.

Explicit formulas are as follows.

(1) Maximum likelihood estimator $\hat{T}_{MLE}$:

solution of the equation

$$N = T_{MLE}[1 - e^{-n/T_{MLE}}]$$

(see Lewontin and Prout (1956) or Huang and Weir (2001));

(2) Turing Good $\hat{T}_{TG}$ :

$$\hat{T}_{TG} = \frac{N}{\hat{C}_{TG}} = \frac{N}{1 - \frac{n_1}{n}};$$

(3) Jackknife estimator $\hat{T}_{J,k}$ of order $k$:

$$T_{J,k} = N + \sum_{j=1}^{k} (-1)^{j+1} \binom{k}{j} n_j$$

(see Burnahm and Overton (1979))

In Burnahm and Overton (1979) there is also a method suggested to determine the most suitable $k$, based on a statistical test that subsequently rejects the various $k$, until the first which is not rejected at some level $\alpha$. We indicate also the estimation results when the "optimal" value of $k$ is selected in that way. The resulting estimator for $\alpha = 0.14$ is denoted by $T_{J,opt}$ and listed in tables 1 and 2.

The low value of $\alpha$ is selected to avoid that all values of $k$ are rejected. All $k$ happen, in fact, to be rejected for the Shakespeare data, so that the method in Burnahm and Overton (1979) does not indicate which $k$ to use.

(4) First estimator of Chao: $\hat{T}_{C_1}$:

$$T_{C_1} = N + n_1^2/(2n_2)$$

(see Chao (1984));

(5) Bias corrected estimator of Chao: $\hat{T}_{C_2}$:

$$T_{C_1} = N + n_1^2/(2(n_2+1)) - (n_1 n_2)/(2(n_2+1)^2)$$

(see Colwell (2006))

(6) First estimator of Chao and Lee: $\hat{T}_{CL_1}$:

$$\begin{aligned} \hat{T}_{CL}(\hat{\gamma}^2) &= \frac{N}{\hat{C}_{TG}} + \frac{n(1-\hat{C}_{TG})}{\hat{C}_{TG}}\hat{\gamma}^2 \\ &= \frac{nN}{n-n_1} + \frac{nn_1}{n-n_1}\hat{\gamma}^2, \end{aligned}$$

with $\hat{\gamma}^2$ given by

$$\hat{\gamma}^2 = \max\left((nN/(n-n_1)\sum j(j-1)n_j/(n(n-1)) - 1, 0\right)$$

(see Chao and Lee (1992));

(7) Second estimator of Chao and Lee $\hat{T}_{CL_2}$:

as above with $\hat{\gamma}^2$ replaced by $\tilde{\gamma}^2$ given by

$$\tilde{\gamma}^2 = \max\left(\hat{\gamma}^2\left((1+n_1\sum j(j-1)n_j/((n-1)(n-n_1))\right), 0\right).$$

Each set of data has seven samples and Table 1 reports the RMS error of the estimators from the true value of the number of taxi cabs (roughly 420 on the days of sampling), which is to say the accuracy of the estimates. One can see that the MLE and $\hat{T}_{TG}$ are among the best estimators in the A data set, but perform rather poorly in the B data set. Other estimators are performing with mixed results, but some Jackknife are quite accurate, and accuracy is often improved when a choice is made on the value of $k$. Pooling all the data (last column), the Jackknife estimator with the optimal choice for $k$ turns out to be the most accuraty among all previously known estimators. Finally, our $\hat{T}$, the regular Bayesian, is performing extremely badly from all points of view, while our $\hat{T}'$ is doing well in the uniform distributions, and extremely well in the non-uniform ones, yielding also the best overall estimator.

Table 2 shows the SE of the estimators, which is the deviation from the average estimated value, indicating the precision of the estimation. Once again, the optimal Jackknife and our $\hat{T}'$ are the most precise, with the first performing slightly better.

An explicit value for the total number of species is also in Edwards and Eberhardt (1967), and it is further discussed in Burnahm and Overton (1979). Here, capture frequencies of a known population of $T = 135$ rabbits are reported; in a sample of $n = 76$ captures, the following are the number of animals captured 1 through 7 times respectively: $(n_1, \dots, n_7) = (43, 16, 8, 6, 0, 2, 1)$. In table 1 we report the performance of the various estimators. Notice that the closest guess is provided by the first Chao estimator, but that several other, including our $\hat{T}'$, come very close to the true value. It is curious that without smoothing and rounding our $\hat{T}'$ would give 135.01.

Next, we present some confidence intervals based on the methods of the present paper, again computed on the Carothers data. The first confidence interval reported is generated by the standard Bayesian method $[\hat{T}(\hat{U} - d), \hat{T}(\hat{U} + d)]$, with $\hat{T}(\hat{U})$ as in (**??**) and $d$ such that $\Psi_{\hat{T}}(\hat{U}_{TG} + d) - \Psi_{\hat{T}}(\hat{U}_{TG} + d) = 1 - \alpha$, where $\alpha$ is the confidence level and $\Psi_{\hat{T}}(\epsilon)$ is as in Lemma (3.1).

The second confidence interval is computed with the mixed method described at the end of the last section. With $d$ computed as just mentioned, the confidence interval is $[\hat{T}'(\hat{U}_{TG} - 2d), \hat{T}'(\hat{U}_{TG} + 2d)]$ with $\hat{T}'(\hat{U})$ as in (**??**).

Table 3 compares our results with those of Chao and with a Jackknife estimate, following the table included in Chao (1984). Note that, as mentioned, except for ours, the other confidence intervals are based on resampling, so each implementation would give different

results. Observe that the confidence intervals based on the Jackknife procedure are narrow, but miss the true value most of the time: this is a feature due to the choice of the order $k$ of the estimator; Chao (1984) uses order $k = 2$, which, from our Table 1, is a bad choice; had the optimal $k$, or at least $k = 4$, been used, then the confidence intervals would have missed the true value much less frequently, with about the same average size. In any case, Chao's confidence intervals miss only once the true value, but at the price of being $5 - 6$ times larger on average. The confindence intervals based on the standard Bayesian estimate are clearly off the mark, but those based on the modified Bayesian estimate (computed by a mixed method with smoothed Turing-Good $\hat{U}_{TG}$ and dilation constant $\bar{c} = 2$), have a much more reasonable coverage, are only twice the size of the Jackknife confidence intervals and do not require resampling.

Table 4 shows the coverage of the true value and the average sizes of 95% and 99% confidence intervals computed using the standard and modified Bayesian methods (again by a mixed method with $\bar{c} = 2$ and a smoothed Turing-Good estimator). Clearly, the 99% confidence interval have a very good coverage (in particular on the less uniform samples) and a moderate interval size.

## Acknowledgements

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.

## References

Almudevar, A., Bhattacharya, R.N. and Sastri, C.C.A. (2000): *Estimating the Probability Mass of Unobserved Support in Random Sampling*, J. Stat. Planning and Inference, **91** , 91-105.

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

Boender, C. G. E. , Rinnoy Kan, A. H. G (1987): *A multinomial Bayesan Approach to the Estimation of Population and Vocabulary Size*, Biometrika **74** No. 4, 849-856.

Böhning, D., Schön, D. (2005): *Nonparametric maximum likelihood estimation of population size based on the counting distribution*, Journal of the Royal Stat. Soc. (C) Appl. Statist. **54**, Part 4, 721-737.

Böhning, D., Suppawattanabe, B., Kusolvisitkul, W., Vivatwongkasem, C (2004): *Estimating the number of drug users in Bangkok 2001: A capture-recapture approach using repeated entries in the list*, Europ. J. of Epidemiology **19**, 1075-1083.

Brose, U., Martinez, M.D., Williams, R. J. (2003): *Estimating species richness: sensitivity to sample coverage and insensitivity to spatial patterns*, Ecology **84** No. 9, 2364-2377.

Bunge, J., Fitzpatrick, M. (1993): *Estimating the number of species: a Review*, J. Amer. Stats. Assn. **88** No. 421, 364-373.

Burnahm, K.P., Overton, W. S. (1979): *Robust estimation of population size when capture probabilities vary among animals*, Ecology **60** No. 5, 927-936.

Carothers, A.D.(1973), *Capture recapture methods applied to a population with known parameters*, J. Animal Ecology, **42**: 125-146.

Chao, A. (1984): *Nonparametric estimation of the number of classes in a population*, Sc. J. of Stat. **11**, 265-270.

Chao, A., Lee, S-M. (1992): *Estimating the number of classes via sample coverage*, J. Amer.Stat.Assn., **87** No.417, 210-217.

Church, K. W. , Gale, W. A. (2006): *Enhanced Good-Turing and Cat-Cal: two new methods for estimating probabilities of english bigrams*, Preprint

Colwell, R.K. (2006), *Estimates*. Software Freeware;
see http://viceroy.eeb.uconn.edu/estimates.

Darroch, J.N., Ratcliff (1980): *A Note on Capture-Recapture Estimation*, Biometrics, **36**, 149-153.

Edwards, W.R, Eberhardt, L.L. (1967): *Estimating cottontail abundance from live trapping data*, J. of Wildlife Manag. **33**, 28-39.

Efron, B. (1981): *Nonparametric standard errors and confidence intervals*, Canadian J. Statist. **9**, 139-172.

Efron, B.,Thisted, R. (1976): *Estimating the number of unseen species: how many words did Shakespeare know?*, Biometrika **63**, 435-467.

Esty, W.W. (1986): The size of a coverage, *Numismatic Chronicle*, **146**, 185-215.

Fisher, R.A., Steven Corbet, A., Williams, C.B. (1943): *The relation between the number of species and the number of individuals in a random sample of an animal population*, J. An. Ecol., **12** No. 1, 42-58.

Gandolfi, A., 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.

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

Harris, B. (1968): *Statistical inference in the classical occupancy problem: unbiased estimation of the number of classes*, J. Amer.Stat.Assn. **63**, 837-847.

Huang, S-P and Weir, B.S. *Estimating the Total Number of Alleles Using a Sample Coverage Method* Genetics 2001 159: 1365-1373.

Huand, J. (2006): *Maximum likelihood estimation of Dirichlet distribution parameters*, Manuscript.

Jedynak, B., Khudanpur, S., Yazgan, A. (2005) *Estimating Probabilities from Small Samples* , 2005 Proceedings of the American Statistical Association, Statistical computing section [CD-ROM], Alexandria, VA : American Statistical Association.

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.

Lehmann, E. L. (1983): *Theory of point estimation*, Wiley ed., New York.

Lewontin, P., Prout, T. (1956): *Estimation of the different classes in a population*, Biometrics **12**, 211-223.

Lijoi, A, Mena, H. R., Prünster, I. (2007) *Bayesian nonparametric estimation of the probability of discovering new species*. Preprint.

Lindsay, B. G., Roeder, K. (1987): *A unified treatment of integer parameter models*, J. Am. Statist. Ass. **82**, 758-764.

Mao, C.X. (2004): *Predicting the conditional probability of discovering a new class*, Journal of the American Statistical Association, **99**, 1108-1118.

Marchand, J.P. and Schroeck, F.E. (1982): *On the Estimation of the Number of Equally Likely Classes in a Population*, Communications in Statistics, Part A–Theory and Methods, **11**, 1139-1146.

McAllester, D. and Schapire, R.E. (2000): *On the Convergence Rate of Good-Turing Estimators*, Conference On Computing Learning Theory (COLT), 1-6.

McNeil, D. (1973): *Estimating an author's vocabulary*, J. Am. Stat. Ass., **68** No. 341, 92-96.

Norris III, J.L., and Pollock, K.H. (1998), *Non-parametric MLE for Poisson species abundance models allowing for heterogeneity between species*, Environ. Ecol. Statist. **5**: 391-402.

Orlitsky, A., Santhanam, N. P, Zhang, J. (2003): *Always Good Turing: Asimptotically Optimal Probability Estimation*, Science, **302** No. 5644, 427-431.

Pitman, J. (2005): *Combinatorial stochastic processes*, Lecture Notes for the St. Flour Summer School.

Shen, T-J., Chao, A., Lin, C-F. (2003): *Predicting the number of new species in further taxonomic sampling*, Ecology, **84** No. 3 , 798-804.

Zabell, S. L.(1982): *W. E Johnson's "Sufficientness" Postulate*, The Annals of Statistics, **10 No. 4** , 1090-1099.

| | | | CAROTHERS DATA | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| DATA | Estimate of number of words known to Shakespeare | Cott. rabbit data | A$\alpha$-RMS Error | A$\beta$-RMS Error | A$\gamma$-RMS Error | B$\alpha$-RMS Error | B$\beta$-RMS Error | B$\gamma$-RMS Error | overall RMS Error |
| $T_{MLE}$ | 31534 | 100.4 | 65.96 | 78.03 | 45.9 | 130.84 | 79.14 | 109.7 | 89.41 |
| $T_{TG}$ | 32054 | 109.1 | 66.06 | 74.86 | 52.13 | 128.65 | 83.97 | 110.61 | 89.93 |
| $T_{JK1}$ | 45910 | 119 | 210.42 | 204.43 | 117.12 | 217.12 | 221.21 | 141.83 | 189.73 |
| $T_{JK2}$ | 55943 | 146 | 145.46 | 137.08 | 67.91 | 150.2 | 159.21 | 83.86 | 128.78 |
| $T_{JK3}$ | 63925 | 165 | 104.8 | 99.05 | 59.45 | 101.31 | 115.66 | 54.27 | 92.12 |
| $T_{JK4}$ | 70685 | 178 | 86.41 | 90.77 | 72.51 | 68.86 | 87.32 | 55.09 | 77.85 |
| $T_{JK5}$ | 76632 | 181 | 84.24 | 100.94 | 107.65 | 58.23 | 76.78 | 90.61 | 87.917 |
| $T_{JK6}$ | 81961 | 162 | 90.57 | 116.89 | 174.73 | 78.44 | 91.8 | 160.65 | 124.36 |
| $T_{JKopt}$ | 81961 | 146 | 80.22 | 85.27 | 43.35 | 73.86 | 85.79 | 57.7 | 72.72 |
| $T_{C1}$ | 55327 | 134 | 72.29 | 79.47 | 64.85 | 124.95 | 97.89 | 110.91 | 94.20 |
| $T_{C2}$ | 55323 | 131 | 74.01 | 66.77 | 67.14 | 110.63 | 105.8 | 113.17 | 91.37 |
| $T_{CL1}$ | 32054 | 118.5 | 66.06 | 83.1 | 52.13 | 129.58 | 79.13 | 109.93 | 90.49 |
| $T_{CL2}$ | 32054 | 115.2 | 66.06 | 83.21 | 52.13 | 129.42 | 79.54 | 110.22 | 90.58 |
| $\hat{T}$ | 46668 | 178 | 286.02 | 376.76 | 198.17 | 330.89 | 151.37 | 78.62 | 258.76 |
| $\hat{T}'$ | 40903 | 140 | 65.15 | 99 | 47 | 90.77 | 58.68 | 53.12 | 71.58 |
| $T_{ACE}$ | 39448 | 128 | | | | | | | |
| $Efron-Thisted$ | 66534 | | | | | | | | |

TABLE 1. Point estimations in Shakespeare and cottontail data and RMS error in estimations from the Carothers Data

| Standard Errors | A$\alpha$-SE Error | A$\beta$-SE Error | A$\gamma$-SE Error | B$\alpha$-RMS Error | B$\beta$-SE Error | B$\gamma$-SE Error | overall SE Error |
|---|---|---|---|---|---|---|---|
| $T_{MLE}$ | 65.94 | 50.07 | 27.98 | 130.38 | 31.02 | 10.49 | 86.05 |
| $T_{TG}$ | 65.61 | 52.21 | 30.5 | 128.56 | 24.78 | 10.37 | 84.67 |
| $T_{JK1}$ | 63.64 | 67.81 | 68.3 | 52.96 | 53.81 | 44.38 | 73.93 |
| $T_{JK2}$ | 77.2 | 83.32 | 64.3 | 58.98 | 60.47 | 31.17 | 77.47 |
| $T_{JK3}$ | 82.88 | 89.84 | 57.45 | 55.78 | 57.37 | 18.56 | 74.32 |
| $T_{JK4}$ | 83.89 | 90.13 | 65.55 | 48.82 | 49.19 | 38.19 | 74.30 |
| $T_{JK5}$ | 83.73 | 87.36 | 100.4 | 50.24 | 46.63 | 84.61 | 87.82 |
| $T_{JK6}$ | 86.13 | 85.22 | 166.82 | 73.64 | 63.81 | 159.52 | 124.10 |
| $T_{JKopt}$ | 78.51 | 74.37 | 43.33 | 70.05 | 31.36 | 11.38 | 69.25 |
| $T_{C1}$ | 66.07 | 67.22 | 37.83 | 124.4 | 17.04 | 14.34 | 83.84 |
| $T_{C2}$ | 61.84 | 61.59 | 38.0 | 106.88 | 18.21 | 13.8 | 75.40 |
| $T_{CL1}$ | 65.61 | 57.02 | 30.5 | 128.93 | 20.16 | 10.18 | 86.33 |
| $T_{CL2}$ | 65.61 | 57.1 | 30.5 | 128.84 | 20.33 | 10.23 | 86.37 |
| $\hat{T}'$ | 100.9 | 90 | 55.30 | 182.25 | 46.839 | 38.17 | 138,14 |
| $\hat{T}'$ | 60.21 | 60.55 | 38 | 87.43 | 37.55 | 10.82 | 79.83 |

TABLE 2. Table of Standard Errors from the Carothers data

| Data subsets | Jackknife estimate | 95% conf. interval | Chao's estimate | 95% conf. interval | Standard Bayesian estimate | 95% conf. interval | Modified Bayesian estimate | 95% conf. interval |
|---|---|---|---|---|---|---|---|---|
| $A\alpha$ a | 192 | 155-229 | 253 | 147-475 | 477 | 413-560 | 304 | 247-408 |
| $A\alpha$ b | 217 | 176-258 | 414 | 230-885 | 734 | 637-861 | 440 | 358-599 |
| $A\alpha$ c | 223 | 182-264 | 484 | 247-1207 | 836 | 726-981 | 493 | 400-674 |
| $A\alpha$ d | 325 | 274-376 | 384 | 251-540 | 676 | 606-759 | 450 | 384-550 |
| $A\alpha$ e | 332 | 281-383 | 366 | 250-513 | 678 | 610-760 | 456 | 390-554 |
| $A\alpha$ f | 350 | 297-403 | 430 | 275-616 | 734 | 660-821 | 487 | 417-592 |
| $A\alpha$ g | 407 | 350-464 | 404 | 283-495 | 679 | 619-747 | 482 | 422-563 |
| Coverage | | 1 in 7 | | 7 in 7 | | 1 in 7 | | 5 in 7 |
| Ave. size c.i. | | 94.57 | | 435.43 | | 174 | | 188.86 |
| $B\alpha$ a | 233 | 190-276 | 691 | 344-1808 | 1109 | 963-1302 | 631 | 512-873 |
| $B\alpha$ b | 199 | 160-238 | 325 | 183-726 | 590 | 510-695 | 362 | 293-494 |
| $B\alpha$ c | 213 | 172-254 | 439 | 226-1123 | 644 | 558-756 | 390 | 318-533 |
| $B\alpha$ d | 333 | 282-384 | 421 | 272-633 | 782 | 701-880 | 510 | 431-622 |
| $B\alpha$ e | 315 | 266-364 | 338 | 227-471 | 610 | 548-684 | 415 | 355-505 |
| $B\alpha$ f | 303 | 250-356 | 331 | 216-465 | 592 | 530-6668 | 402 | 342-491 |
| $B\alpha$ f | 346 | 307-385 | 312 | 224-380 | 546 | 497-603 | 399 | 348-468 |
| Coverage | | 0 | | 6 in 7 | | 0 in 7 | | 5 in 7 |
| Ave. size c.i. | | 90 | | 559.14 | | 182.71 | | 198.14 |

TABLE 3. Comparison of confidence intervals

| | Standard Bayesian Estimate | | | | Modified Bayesian Estimate | | | |
|---|---|---|---|---|---|---|---|---|
| Data sub-sets | Coverage 95% c.i. | Ave. size 95% c.i. | Coverage 99% c.i. | Ave. size 99% c.i. | Coverage 95% c.i. | Ave. size 95% c.i. | Coverage 99% c.i. | Ave. size 99% c.i. |
| $A\alpha$ | 1 in 7 | 174 | 1 in 7 | 231.14 | 5 in 7 | 188.86 | 7 in 7 | 257.71 |
| $A\beta$ | 0 | 197.71 | 0 | 262.57 | 5 in 7 | 213.43 | 5 in 7 | 291.57 |
| $A\gamma$ | 0 | 112.28 | 0 | 148.14 | 4 in 7 | 126.71 | 6 in 7 | 168.71 |
| $B\alpha$ | 0 | 182.71 | 0 | 242.86 | 5 in 7 | 198.14 | 6 in 7 | 271.71 |
| $B\beta$ | 1 in 7 | 144 | 1 in 7 | 191 | 6 in 7 | 158.14 | 7 in 7 | 215 |
| $B\gamma$ | 1 in 7 | 92.43 | in 7 | 122.14 | 4 in 7 | 105.86 | 7 in 7 | 140.57 |
| Total | 3 in 42 | 150.52 | 4 in 42 | 199.64 | 29 in 42 | 165.19 | 38 in 42 | 224.21 |

TABLE 4. Evaluation of confidence intervals between Standard Bayesian and Modified Bayesian methods for the Carothers data.