Seen once or more than once: applying Good–Turing theory to estimate species richness using only unique observations and a species list

Anne Chao\*1, Robert K. Colwell\textsuperscript{2,3,4}, Chun-Huo Chiu\textsuperscript{1} and Ditch Townsend\textsuperscript{5}

\textsuperscript{1}Institute of Statistics, National Tsing Hua University, Hsin-Chu 30043, Taiwan; \textsuperscript{2}Department of Ecology and Evolutionary Biology, University of Connecticut, Storrs, CT 06269, USA; \textsuperscript{3}University of Colorado Museum of Natural History, Boulder, CO 80309, USA; \textsuperscript{4}Departamento de Ecologia, Universidade Federal de Goiás, CP 131, 74.001-970 Goiânia, GO, Brasil; and \textsuperscript{5}4 Park Cottages, Chawleigh, Devon EX18 7LB, UK

Summary

1. Due to sampling limitations, almost every biodiversity survey misses species that are present, but not detected, so that empirical species counts underestimate species richness. A wide range of species richness estimators has been proposed in the literature to reduce undersampling bias. We focus on nonparametric estimators, which make no assumptions about the mathematical form of the underlying species abundance/incidence distributions.

2. For replicated incidence data, in which only species presence/absence (or detection/non-detection) is recorded in multiple sampling units, most existing nonparametric estimators of the number of undetected species are based on the frequency counts of the uniques (species detected in only one sampling unit) and duplicates (species detected in exactly two sampling units).

3. Some survey methods, however, record only uniques and super-duplicates (species observed in more than one sampling unit). Using the Good–Turing frequency formula, we developed a method to estimate the number of duplicates for such data, allowing estimation of true species richness, including undetected species.

4. We test our estimators on several empirical datasets for which doubletons were recorded and on simulated sampling data, then apply them to an extensive, but previously unusable survey of coral reef fishes, for which only uniques and super-duplicates were recorded.

5. We extend the method to abundance data and discuss other potential applications.

Key-words: abundance data, citizen science, coral reef fishes, Good–Turing frequency formula, incidence data, singletons, species richness, super-doubletons, super-duplicates, uniques

Introduction

In biodiversity and conservation studies, species richness plays a primary role in characterizing communities, assemblages and habitats (Magurran 2004; Magurran & McGill 2011). As a simple count, species richness has intuitive mathematical properties and features prominently in foundational models of community ecology. Comparisons of species richness among multiple assemblages help researchers to understand the processes that produce and maintain biodiversity and assess the effect of human activities on biodiversity. However, in species-rich ecosystems, such as coral reefs and rain forests, inventorying every species is impractical, or takes so much time and effort that comparisons remain out of reach.

Based on incomplete sampling data, two statistical approaches have been used to estimate species richness (observed plus undetected): parametric and nonparametric methods. Parametric approaches require specific parametric assumptions about the species abundance distribution. However, different parametric distributions may fit the observed data equally well, but lead to drastically different species richness estimates. A practical problem is that extensive numerical procedures are typically required to find the species richness estimates under parametric models, and the iterative steps sometimes fail to converge properly; therefore, species richness estimates may not be obtainable. Nonparametric richness estimators that are always obtainable and make no assumptions about the mathematical form of the underlying species abundance/incidence distributions have thus earned a place in such studies (Colwell & Coddington 1994; Gotelli & Colwell 2011; Chao & Chiu 2012; Gotelli & Chao 2013).

For both practical and biological reasons, recording the presence/absence or detection/non-detection of species in replicated sampling units (incidence data) is often preferable to enumerating individuals (abundance data). For microbes, colonial plants and sessile invertebrates, individuals are difficult or impossible to define. For mobile organisms, replicated incidence data are less likely to double-count individuals. For social animals, counting the individuals in a colony, flock,
or school may be difficult or impractical. Perhaps surprisingly, replicated incidence data support statistical approaches to richness estimation that are just as powerful as corresponding abundance-based approaches (Chao et al. 2014; Chao & Colwell 2017). Moreover, replicated incidence data account for spatial (or temporal) heterogeneity in the data, whereas abundance-based approaches often cannot (Colwell, Mao & Chang 2004; Colwell et al. 2012) – an important additional advantage. Although our model formulation in the main text of this study is focused on incidence data, the same approach can be used to deal with abundance (enumeration) data; see Discussion.

A counter-intuitive, but fundamental concept in estimating the number of undetected species for incidence data is that frequent species (which are likely to be detected in most sampling units) contain almost no information about the richness of undetected species. By contrast, infrequent species (which are likely to be either undetected or detected in only a few sampling units) contain nearly all the information about undetected species richness (Gotelli & Chao 2013). Consequently, most nonparametric estimators of the number of undetected species, e.g. the Chao2 estimator (Chao 1987) and the jackknife estimators of the first two orders (Burnham & Overton 1978), are based on the information about the infrequently detected (rare) species. To date, the minimum information required for the simplest incidence-based richness estimators has been (i) a count of the total number of species observed, (ii) the number of species observed in only one sampling unit (uniques) and (iii) the number of species observed in exactly two sampling units (duplicates). The information from uniques and duplicates is used to estimate the number of undetected species.

In this study, we introduce a novel approach to estimate species richness that requires only the total number of species observed and the number of uniques. It does not require a count of duplicates at all. Instead, we estimate the true number of duplicates from the observed numbers of uniques and super-duplicates (species observed in at least two sampling units). Then the observed number of uniques and the estimated number of duplicates are applied to infer the number of undetected species and total species richness using the classic Chao2 formula (Chao 1987). As proved by Chao (1987), the Chao2 estimator provides a nonparametric lower bound of species richness. Recently, A. Chao, C.H. Chiu, R.K. Colwell, R.L. Chazdon & N.J. Gotelli (under revision) further showed that the Chao2 estimator is an unbiased point estimator as long as very rare species (specifically, undetected species and unique species in samples) have approximately equal detection probability; the other species (those detected in two or more samples) could be highly heterogeneous without affecting the estimator. On the other hand, if very rare species are heterogeneous and the sample size is not sufficiently large, at best we can provide only a lower bound, because in such cases the data available do not contain sufficient information to accurately estimate species richness.

Our estimation of the true number of duplicates is based on the Good–Turing frequency formula, which was originally developed by Alan Turing and his colleague I. J. Good for their work during World War II in breaking the German ‘Enigma’ code (Good 1953, 2000). They used the formula to estimate the true relative frequencies of rare code elements (including undetected elements), based on observed frequencies in intercepted ‘samples’ of code. In our derivation, we have adapted the same approach to estimate the true count of duplicates, making possible estimation of the number of undetected species and, thus, true species richness. We first test our estimators on several empirical datasets with a known number of duplicates. Second, we simulate sampling from hypothetical assemblages with a broad spectrum of relative incidence patterns to investigate the general performance of our method. Third, we demonstrate the utility of our method to estimate richness for a unique, extensive survey of coral reef fishes that lacks a known number of duplicates.

**Empirical datasets**

The original motivation for this study was an extensive survey of coral reef fishes conducted in Tunku Abdul Rahman Marine Park in the South China Sea, off Sabah in Malaysian Borneo (Townsend 2015). Between 2006 and 2009, this citizen science project accumulated observations from 116 SCUBA diving and snorkelling sessions (sampling units), each 5 to 97 min in duration, totalling more than 95 h (49 min on average). In each session, every newly encountered species or colour morph was photographed (Fig. 1), and subsequently photographed again if encountered at a later date (i.e. in a different sampling unit). Third and later encounters were not photographed (except in uncertain cases), so there is no way to distinguish true duplicates (species photographed twice, but never seen again) from species seen more than twice. In all, several thousand fish photographs were amassed. Hence, the data allow certainty only regarding uniques (101 species) (Fig. 1) and species seen more than once (super-duplicates, 340 species), but offers no information on true duplicates. The number of observed species was 441: the sum of the counts of uniques and super-duplicates. An incidence-based approach is especially advantageous for social species (for which individuals are difficult to distinguish and to count), and avoids the risk of double-counting individuals within sessions. To reduce the chance of recording the same individuals twice, an effort was made to avoid visiting the same site in consecutive sessions.

In addition, we apply our method to several empirical datasets for which the true number of duplicates in the data is known. This allows us to compare the estimated number of duplicates, based on our new estimator, with the observed number. The following 15 datasets are based on six sources/studies. Data in each set consist of a species-by-sampling unit incidence (presence/absence or detection/non-detection) matrix. The data required for our analysis, for each set, are provided in Table 1.

1. Because ants are colonial organisms, incidence data, which tend to reflect the number of colonies, are much more appropriate than abundance data. We first considered five ant datasets from Longino & Colwell (2011), who collected ant
Species richness estimation

Fig. 1. Examples of reef fish species seen only once (uniques) in a survey of coral reef fishes conducted in Tunku Abdul Rahman Marine Park in the South China Sea, in Malaysian Borneo. Based on the number of uniques (101) and the number of species seen twice or more (super-duplicates, 340), we used a novel method to estimate the total number of fish species in the assemblage (543), including about 102 species not detected in the survey. Pictured here are (left-to-right, top-to-bottom) Uranoscopus bicinctus (Marbled Stargazer), Chrysiptera springeri (Springer’s Damselﬁsh), Pleurosigia boldinghi (Soft Coral Ghostgoby), Pomacanthus semicirculatus (Semicircle Angelfish), Calloplepis altivelis (Comet) and Apogonichthyoides timorensis (Timor Cardinalﬁsh).

Table 1. Numerical test results of the proposed estimator of the count of duplicates for several datasets

<table>
<thead>
<tr>
<th>Data</th>
<th>$T$</th>
<th>$S_{obs}$</th>
<th>CV (mean)</th>
<th>Observed $Q_1$</th>
<th>Observed $Q_2$</th>
<th>Estimated $Q_2$</th>
<th>Chao2 with observed $Q_2$ (95% CI)</th>
<th>Chao2 with estimated $Q_2$ (95% bootstrap CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ants (50 m)</td>
<td>599</td>
<td>227</td>
<td>2.11 (0.036)</td>
<td>49</td>
<td>23</td>
<td>24.3</td>
<td>279.1 (254.6, 325.3)</td>
<td>276.4 (259.3, 302.7)</td>
</tr>
<tr>
<td>(500 m)</td>
<td>230</td>
<td>241</td>
<td>1.95 (0.041)</td>
<td>71</td>
<td>34</td>
<td>34.1</td>
<td>314.8 (284.3, 366.7)</td>
<td>314.5 (294.2, 342.3)</td>
</tr>
<tr>
<td>(1070 m)</td>
<td>150</td>
<td>122</td>
<td>1.57 (0.079)</td>
<td>28</td>
<td>16</td>
<td>13.8</td>
<td>146.3 (132.5, 178.7)</td>
<td>150.2 (139.1, 168.9)</td>
</tr>
<tr>
<td>(1500 m)</td>
<td>200</td>
<td>56</td>
<td>1.75 (0.076)</td>
<td>13</td>
<td>4</td>
<td>6.41</td>
<td>77.0 (62.9, 119.8)</td>
<td>69.1 (62.6, 81.8)</td>
</tr>
<tr>
<td>(2000 m)</td>
<td>200</td>
<td>14</td>
<td>1.18 (0.095)</td>
<td>1</td>
<td>2</td>
<td>0.50</td>
<td>14.2 (14.0, 19.2)</td>
<td>15.0 (14.3, 17.0)</td>
</tr>
<tr>
<td>Ants (Malaise)</td>
<td>62</td>
<td>103</td>
<td>1.49 (0.051)</td>
<td>39</td>
<td>18</td>
<td>17.7</td>
<td>144.6 (123.5, 187.2)</td>
<td>145.2 (129.9, 168.9)</td>
</tr>
<tr>
<td>(Fogging)</td>
<td>459</td>
<td>165</td>
<td>1.77 (0.039)</td>
<td>21</td>
<td>14</td>
<td>10.5</td>
<td>180.7 (170.8, 207.0)</td>
<td>185.9 (177.5, 199.8)</td>
</tr>
<tr>
<td>(Berlese)</td>
<td>217</td>
<td>117</td>
<td>1.75 (0.025)</td>
<td>33</td>
<td>19</td>
<td>16.0</td>
<td>145.5 (130.0, 179.5)</td>
<td>150.9 (137.2, 174.2)</td>
</tr>
<tr>
<td>Ciliates (Central)</td>
<td>17</td>
<td>130</td>
<td>1.54 (0.111)</td>
<td>63</td>
<td>28</td>
<td>25.7</td>
<td>196.7 (168.0, 247.0)</td>
<td>202.7 (179.4, 236.9)</td>
</tr>
<tr>
<td>(Etosha Pan)</td>
<td>19</td>
<td>216</td>
<td>1.19 (0.078)</td>
<td>107</td>
<td>44</td>
<td>42.9</td>
<td>339.3 (290.2, 405.5)</td>
<td>342.3 (310.9, 383.8)</td>
</tr>
<tr>
<td>(Southern)</td>
<td>15</td>
<td>130</td>
<td>1.48 (0.088)</td>
<td>82</td>
<td>28</td>
<td>30.5</td>
<td>262.1 (219.3, 331.1)</td>
<td>258.2 (221.8, 297.1)</td>
</tr>
<tr>
<td>Birds (2015)</td>
<td>16</td>
<td>223</td>
<td>0.74 (0.098)</td>
<td>31</td>
<td>20</td>
<td>17.9</td>
<td>245.5 (232.9, 274.4)</td>
<td>248.2 (237.4, 266.8)</td>
</tr>
<tr>
<td>(2016)</td>
<td>17</td>
<td>223</td>
<td>0.68 (0.539)</td>
<td>23</td>
<td>14</td>
<td>13.6</td>
<td>240.8 (230.0, 268.4)</td>
<td>241.3 (233.3, 255.3)</td>
</tr>
<tr>
<td>Tree seedlings</td>
<td>121</td>
<td>34</td>
<td>1.50 (0.225)</td>
<td>3</td>
<td>2</td>
<td>1.5</td>
<td>36.2 (34.3, 31.4)</td>
<td>36.9 (34.6, 47.8)</td>
</tr>
<tr>
<td>Cottontail rabbit</td>
<td>18</td>
<td>76</td>
<td>0.89 (0.059)</td>
<td>43</td>
<td>16</td>
<td>15.5</td>
<td>130.6 (104.2, 181.6)</td>
<td>132.4 (108.1, 174.7)</td>
</tr>
</tbody>
</table>

$T$, number of sampling units; $S_{obs}$, observed number of species; mean and CV, the estimated mean and CV (coefficient of variation) of species detection probabilities in each sampling unit; observed $Q_1$, the observed true count of uniques; observed $Q_2$, the observed true count of duplicates; estimated $Q_2$, the estimated count of duplicates, obtained by solving eqn (7); 95% CI for Chao2 with observed $Q_2$, based on the Chao (1987) asymptotic variance formula via a log transformation; 95% CI for Chao2 with estimated $Q_2$, based on a bootstrap method with 200 replications (Appendix S2) via a log transformation. All results were obtained by means of the R code in Appendix S3 or the online software ‘SUPERDUPLICATES’.

data from mini-Winkler samples (as replicated sampling units) for five elevations on the Barva Transect, a 30-km continuous gradient of wet forest on Costa Rica’s Atlantic slope. The five sites were, respectively, at elevations 50, 500, 1070, 1500 and 2000 m; there were, respectively, 599, 230, 150, 200 and 200 sampling units. Species presence or absence was recorded in each sampling unit, which consisted of all worker ants extracted from a 1-m² forest floor plot. The incidence frequency counts appear in Colwell et al. (2012), their table 6.

2. A second group of datasets for tropical rain forest ants comes from Longino, Coddington & Colwell (2002). Ants were captured in Costa Rica using three sampling techniques: (i) Berlese extraction of soil samples (217 samples), (ii) fogging samples from canopy fogging (459 samples) and (iii) Malaise trap samples for flying and crawling insects (62 samples). Each sampling technique represents a different assemblage. Species presence/absence was recorded in each capture sample.

3. The third group of datasets includes soil ciliate incidence data for a total of 51 soil samples from three areas of Namibia, Africa: Etosha Pan (19 samples), Central Namib Desert (17 samples) and Southern Namib Desert (15 samples). As it was not feasible to count the exact numbers of ciliates in a soil sample, only species detection/non-detection in each sample was recorded. The original data are provided in Foissner, Agatha & Berger (2002), p.58–63.

4. The fourth group of datasets was taken from the records of the Hong Kong Big Bird Race (BBR). The Hong Kong BBR is an annual competition among teams of birdwatchers. The challenge is to record as many bird species as possible during a fixed interval of time in the Hong Kong territory. In 2015, 16 teams competed, whereas 17 teams competed in 2016. The raw data for each year consist of bird detection/non-detection records for each team; see http://www.wwf.org.hk/en/your_support/join_our_events/hkbbr/; then click on Big Bird Race/Event History for data sources.

5. The fifth dataset we considered was the seed-bank data provided by Butler & Chazdon (1998). They recorded the species of tree seedlings that germinated from each of 121 soil samples (each soil sample is regarded as a sampling unit). The original data consist of species-by-sample abundance records. Here, we convert the abundance matrix into an incidence matrix. That is, a ‘presence’ of species means that at least one seedling was counted and the exact abundance is ignored. The corresponding analysis based on species abundances will be discussed later.

6. The last dataset was based on the cottontail rabbit capture-recapture data provided by Edwards & Eberhardt (1967). They conducted a live-trapping study on a confined cottontail population of known size. In their study, 135 cottontail rabbits were penned in a 4-acre rabbit-proof enclosure. Live trapping was conducted for 18 consecutive nights. A unique tag/mark was attached to each animal to distinguish individuals. For this example, each individual cottontail can be regarded as a ‘species’ so that the Chao2 formula can be directly applied to infer population size. The data have the advantage of known population size (135) so that the performance of estimators can be examined.

Model and estimation theory

REPLICATED INCIDENCE DATA

We extend the incidence-based sampling model of Colwell et al. (2012) to encompass datasets for which only the numbers of uniques and super-duplicates are known. Assume that there are S species in the target assemblage and T sampling units (in the reef fish example, each sampling unit corresponds to a ‘diving or snorkelling session’; in the ant examples, each mini-Winkler extraction or Berlese/fogging/Malaise sample is a sampling unit; in the ciliate and seed-bank data, each soil sample is a sampling unit; in the bird and cottontail data, each team/trapping-night is a sampling unit). The detection or non-detection of each species within each sampling unit is recorded to form a species-by-sampling-unit incidence matrix \(W_j\) with \(S\) rows and \(T\) columns. The value of the element \(w_{ij}\) of this matrix is 1 if species \(i\) was detected in the \(j\)-th sampling unit, and 0 if species \(i\) was not detected in the \(j\)-th sampling unit. The row sum of the incidence matrix \(Y_j = \sum_{i=1}^{S} w_{ij}\) denotes the incidence-based frequency of species \(i, j = 1, 2, \ldots, S\).

We assume that the detection probability of the \(i\)-th species in the \(j\)-th sampling unit is affected by two sources of heterogeneity: a species effect and sampling-unit effect. The species effect accounts for the fact that each species may have its own unique incidence rate \(\pi_i\); the sampling-unit effect \(v_j\) recognizes that species detection in the \(j\)-th sampling unit may be affected by some known or unknown factors specific to that sampling unit. In the reef fish example, a sampling-unit effect may arise from different dive duration, site, depth, visibility, season and weather variability, instruments (camera, lens), an observer’s capability/characteristics, etc. In the first ant example, the soil and litter characteristics or weather conditions may vary among the sampling sites for the mini-Winkler leaf-litter extractions; similar interpretations can be made for other datasets.

Our model assumes a multiplicative form for the species effect and sampling-unit effect; that is, the detection probability of the \(i\)-th species in the \(j\)-th session is expressed as \(\pi_i v_j\). This kind of model has been widely used in the context of capture-recapture research to estimate the size of a population; it is referred to as Model \(M_{h0}\), where the sub-index \(h\) denotes heterogeneity among individual capture probabilities and the sub-index \(t\) denotes time-varying effects (Chao 2001). Here, we adapt this model to estimate species richness based on incidence data. As many factors are involved in the sampling-unit effects, \(\{v_1, v_2, \ldots, v_T\}\), we assume that they can be regarded as random variables taken from an unknown probability density function \(h(v)\). Here, we regard the species effects \(\{\pi_1, \pi_2, \ldots, \pi_S\}\) as fixed parameters, but treat the sampling-unit effects \(\{v_1, v_2, \ldots, v_T\}\) as random variables. The conditional probability distribution of \(W_j\) given \(v_j\) is a Bernoulli random variable with probability of success \(\pi_i v_j\). That is,

\[
P(W_j = w_{ij} \mid v_j) = (\pi_i v_j)^{w_{ij}} (1 - \pi_i v_j)^{1-w_{ij}},
\]

\[i = 1, \ldots, S, j = 1, \ldots, T.\]
This leads to the following unconditional binomial model for the incidence-based frequency $Y_i$:

$$P(Y_i = y_i) = \binom{T}{y_i} \pi_i \int vh(v)dv \left[1 - \pi_i \int vh(v)dv\right]^{T-y_i}$$

$$= \binom{T}{y_i} \lambda_i \left(1 - \lambda_i\right)^{T-y_i}, \quad \text{eqn 1}$$

where $\lambda_i = \pi_i \int vh(v)dv$. That is, the frequency $Y_i$ is a binomial random variable with detection probability $\lambda_i$.

Let $Q_k$ denote the incidence-based frequency counts, i.e. the number of species that are detected in exactly $k$ sampling units, $k = 0, 1, \cdots, T$. The number of observed species is $S_{obs} = \sum_{k=1}^{T} Q_k$. Here, $Q_1$ represents the number of unique species (those that are each detected in only one sampling unit), and $Q_2$ represents the number of duplicate species (those that are each detected in exactly two sampling units), and so on.

The statistical inference problem in this study is to estimate species richness based on the information of $Q_2$ and the number of super-duplicates $Q_{2+} = Q_2 + Q_3 + \cdots + Q_T$. In this case, we have $S_{obs} = Q_1 + Q_{2+}$. As we will show below, our inference is quite robust to the number of sampling units $T$, if $T$ is large enough. Based on the binomial model in eqn (1), we have a general form for the expected value of $Q_k$:

$$E(Q_k) = E\left(\sum_{i=1}^{S} I(Y_i = k)\right) = \sum_{i=1}^{S} \binom{T}{k} \lambda_i^k (1 - \lambda_i)^{T-k}, \quad k = 0, 1, \cdots, T, \quad \text{eqn 2}$$

where $RA(A)$ is an indicator function that equals 1 when $A$ is true and 0 otherwise.

As described in the Introduction, most nonparametric estimators of the number of undetected species for incidence data are based on the counts of the detected infrequent species, especially the numbers of unique and duplicates. Chao (1987) derived a lower bound of undetected species richness in terms of the counts of unique and duplicates; the corresponding lower bound of species richness given below is referred to as the Chao2 estimator (Colwell & Coddington 1994),

$$\hat{S}_{Chao2} = \begin{cases} S_{obs} + \left[\frac{(T-1)}{T}\right]Q_1^2/(2Q_2) & \text{if } Q_2 > 0 \\ S_{obs} + \left[\frac{(T-1)}{T}\right]Q_1(Q_1 - 1)/2 & \text{if } Q_2 = 0 \end{cases} \quad \text{eqn 3}$$

Based on the values of $Q_1$ and $Q_{2+}$, we need first to estimate the true count of duplicates, $Q_2$, to apply the above Chao2 richness estimator. Our estimation is based on the Good–Turing frequency formula, which was developed by Alan Turing and I. J. Good. Turing never published the theory but gave permission to Good to publish it. Two influential papers by Good (1953) and Good & Toulmin (1956) presented Turing’s wartime statistical work on the frequency formula and related topics. In an ecological context for abundance data, Good–Turing frequency theory estimates the true mean relative abundance of those species that appeared $r$ times, $r = 0, 1, \cdots$ in a sample of size $n$. Good and Turing derived the frequency formula for the case of small $r$, i.e. rare species (or rare code elements, in Turing’s case). Chao et al. (2014), Appendix H, extended Good–Turing theory to incidence data. Here, we specifically focus on $r = 1$, i.e. uniques for incidence data, as discussed below.

Let $\lambda_{i1}$ denote the mean detection probability of all unique species. Following typical Good–Turing theory, we assume that the detection probabilities of unique species do not vary greatly among such species, so that most of the detection probabilities of uniques are concentrated at the average $\lambda_{i1}$. Then we can express each of the expected frequencies for doubletons and more frequent species, $E(Q_2), E(Q_3), \cdots, E(Q_T)$, in terms of $\lambda_{i1}$ and $E(Q_1)$ as follows:

$$\frac{E(Q_k)}{T} = \left(\frac{k-1}{k}\right) \frac{E(Q_1)}{T}, \quad k = 2, 3, \cdots, T. \quad \text{eqn 4}$$

Details of the approximation leading to the final expression can be found in Appendix S1, Supporting Information. Therefore, to get the expected frequency of super-doubletons, we sum all expressions in (4) over $k = 2, 3, \cdots, T$, leading to

$$E(Q_{2+}) = \frac{E(Q_1)}{T} \left[ \frac{T}{2} \left(\frac{\lambda_{i1}}{1 - \lambda_{i1}}\right)^{T-1} + 2! \left(\frac{T}{3} \left(\frac{\lambda_{i1}}{1 - \lambda_{i1}}\right)^{T-2} + \cdots + (T-1)! \left(\frac{T}{T} \left(\frac{\lambda_{i1}}{1 - \lambda_{i1}}\right)^{T-3} \right) \right) \right]. \quad \text{eqn 5}$$

Chao et al. (2014) obtained, but did not provide the formal derivation for the following estimator of the mean detection probability of uniques, $\lambda_{i1}$:

$$\hat{\lambda}_{i1} = \frac{2Q_2}{(T-1)Q_1 + 2Q_2}. \quad \text{eqn 6}$$

The proof is given in Appendix S1. Substituting eqn (6) into eqn (5) and replacing the expected values by the observed data, we obtain an approximation for the number of super-doubletons, in terms of observed $Q_1$ and unknown $Q_2$,

$$Q_{2+} \approx Q_1 \left(\frac{Q_1}{2Q_2}\right)^2 \left[1 + \frac{2Q_2}{Q_1} + \frac{2Q_2}{Q_1} + \frac{2Q_2}{Q_1} + \frac{2Q_2}{Q_1} + \frac{2Q_2}{Q_1} + \frac{2Q_2}{Q_1} + \frac{2Q_2}{Q_1} \right]. \quad \text{eqn 7}$$

Under the condition $Q_1 > 2Q_2$, successive terms in the above series typically decline in magnitude very quickly. Given observed values for uniques and super-doubletons, $Q_1$ and $Q_{2+}$, and a fixed number of sampling units $T$, there is no analytic solution, and thus a numerical algorithm is needed to find the solution for $Q_2$. The resulting solution, along with $Q_1$, is substituted in eqn (3) to obtain a species richness estimate. We propose a novel bootstrap method in Appendix S2 to assess sampling variance and obtain the associated confidence intervals. The uncertainty due to the estimation of the count of duplicates is taken into account in our bootstrap procedures. The R code supplied in Appendix S3 can be applied to obtain
the numerical solution of eqn (7) for any given \(T, Q_1\) and \(Q_2\), as well as an estimate of true species richness (eqn 3) with the bootstrap SE and 95% confidence intervals. For readers without R background, we also provide the online software ‘SUPERDUPQUENTES’ available from https://choa.shinyapps.io/SuperDupQuentes to facilitate all computations.

When \(T\) is unknown but large enough, we can let \(T\) tend to infinity (the tail sequences have a negligible sum, as will be demonstrated by our data below) and eqn (7) reduces to

\[
Q_2 \approx \frac{Q_1^2}{2Q_2} \left[ -\log(1 - \frac{2Q_2}{Q_1}) - \frac{2Q_2}{Q_1} \right].
\]  

Eqn 8

The solution for unknown \(Q_2\) is unique because we can prove mathematically that the function

\[
h(Q_2) = \frac{Q_1}{2Q_2} \left[ -\log(1 - \frac{2Q_2}{Q_1}) - \frac{2Q_2}{Q_1} \right]
\]

is a strictly increasing function from 0 to infinity, when \(Q_2\) is increased from 0 to any large integer. In this case, calculations become much simpler. All computation can be performed using the R code (Appendix S3) or the online software ‘SUPERDUPQUENTES’ mentioned above. Thus, when \(T\) is unknown but large enough, we can nonetheless obtain a species richness estimate, given \(Q_1\) and \(Q_2\). Here, how large is large enough? That depends on the species incidence distribution. Roughly, with regard to the estimated number of duplicates, the relative error between the two solutions obtained from eqn (7) and eqn (8) is less than 10% if the unknown true value of \(T > 10\), and less than 5% if the unknown true value of \(T > 100\).

ABUNDANCE DATA

The model and estimators, above, are based on incidence data from a set of replicated sampling units. With slight modifications, our approach works equally well with abundance data. Full derivation details are provided in Appendix S4. Here, we highlight the major differences between the analyses for incidence and abundance data. (i) For abundance data, a single sample of individuals is considered, rather than a set of replicated sampling units. (ii) The statistical inference problem for incidence data is to estimate species richness based on the information carried by uniques and super-duplicates. For abundance data, we estimate species richness based on the information carried by singletons (the number of species observed only once in the single sample) and super-doubletons (the number of species observed at least twice in the sample). (iii) For incidence data, our derivation was based on a model in which the sample incidence-based frequency follows a binomial distribution, characterized by the number of sampling units and the detection probability. The detection probability is a product of the species incidence rate (the species effect) and the sampling-unit effect, as specified in eqn (1). For abundance data, the derivation requires that the sample individual-based frequencies follow a binomial distribution characterized by sample size and a detection probability for each species that is the product of species abundance and individual detectability.

The detectability of an individual may vary by species, and is determined by many possible factors (such as colour or size and, for animals, movement patterns, cryptic behaviours and vocalizations). Because both models are based on the binomial distribution, all derivations are parallel.

In Appendix S4, we report on tests of our method using several real-world surveys with abundance data and known doubleton counts. When sample size is large (which is the case for almost all abundance datasets), the doubleton count, and the estimated true richness, along with the bootstrap SE and 95% confidence intervals, can be estimated using the R code (Appendix S3) or the online software ‘SUPERDUPQUENTES’, given counts of singletons and super-doubletons.

Data analysis

We tested our proposed incidence-based estimator of the count of duplicates on several empirical datasets for which the true observed number of duplicates is known. In Table 1, we show the test results for all datasets described earlier in the Empirical Datasets section. For each dataset, Table 1 shows the basic data information including the number of sampling units \(T\), the observed species richness \((S_{\text{obs}})\), the observed number of uniques \((Q_1)\) and the observed number of duplicates \((Q_2)\). We also give the estimated species mean detection probability per sampling unit and the estimated coefficient of variation (CV, the ratio of standard deviation to mean), which quantifies the degree of heterogeneity among species detection probabilities in each sampling unit. CV = 0 means that detection probability is homogeneous among species. The estimated CV values range from 0.63 to 2.11, signifying that heterogeneity is present among detection probabilities in all datasets. The CV estimator usually exhibits negative bias (Chao & Chiu 2012), implying that the true CV in each case is likely to be higher than the estimated CV.

For each dataset, in Table 1 we also show the estimated count of duplicates, obtained by solving eqn (7), and the Chao2 species richness estimates using, respectively, the observed and the estimated duplicates, along with 95% confidence intervals based on a log transformation (Chao 1987). Table 1 reveals that the proposed estimated count of duplicates is generally (11 of 15 datasets) lower than the true count, leading to slightly higher Chao2 estimates than the corresponding estimate based on the true number of duplicates. In one set (ant data at 500-m elevation), there is almost no difference between the estimated and the true counts. In the remaining three cases, the proposed estimated count of duplicates is somewhat higher than the true count, leading to slightly lower Chao2 estimates than the estimate based on the true number of duplicates. These three cases are as follows: ant data at 500-m elevation (CV = 2.11, mean = 0.036), ant data at 1500 m elevation (CV = 1.75, mean = 0.076) and ciliate data in the Southern Namib area (CV = 1.48, mean = 0.088). These three datasets generally have relatively higher CV and lower mean species detection probability per sampling unit than those of other datasets. These results reveal that the direction of the bias depends on species detection probabilities principally through...
their mean and CV values; see the simulation results for the pattern of dependence on sample size. In the cottontail capture-recapture data, the true population size is 135. The Chao2 estimates with the observed and estimated values of $Q_k$ yield very close population sizes of 131 and 133, respectively.

Overall, it seems remarkable that our estimated counts of duplicates so closely match the observed true counts. The Chao2 estimate based on the estimated count of duplicates also differs relatively little from the estimate calculated from the observed true data. The 95% confidence interval based on the Chao2 estimate using the observed counts of duplicates is generally wider than that using the estimated counts (Table 1), probably because the sampling variation in the observed count of duplicates inflates the uncertainty of the resulting Chao2 estimate to a larger extent relative to the increased uncertainty due to the estimation of $Q_k$.

**Simulation Results**

To examine the performance of the estimated count of duplicates (using eqn 7) and the derived species richness estimator based on the information provided by $Q_1$ and $Q_{2+}$, we generated incidence-based data from several hypothetical sets of detection probabilities $(\lambda_1, \lambda_2, \ldots, \lambda_{200})$ underlying our model (eqn 1). In Table 2, we selected eight cases for which the true CV value ranges between 0 and 4.22. In all cases, the true species richness was fixed to be $S = 200$. The eight cases are separated into two groups. In Case 1 through Case 4 (Random Parameters Group), for which the detection probabilities $(\lambda_1, \lambda_2, \ldots, \lambda_{200})$ were determined from random variables generated from a statistical distribution, we first generated a set of 200 random variables and used them consistently through all simulated trials. The mean detection probability and the CV of the generated random variables are also shown in Table 2 for each case. In Cases 5–8 (Fixed Parameters Group), for which $(\lambda_1, \lambda_2, \ldots, \lambda_{200})$ are fixed parameters, we show in Table 2 the true values for the mean and CV of $(\lambda_1, \lambda_2, \ldots, \lambda_{200})$.

In the simulations, the number of sampling units was selected to vary from 4 to 100 in increments of 4. For each fixed case of detection probabilities and fixed number of sampling units, we generated 500 sets of incidence matrices, stochastically, to examine the performance of our estimated count of duplicates and the corresponding species richness estimates.

The simulation results are plotted in Figs 2 and 3. For each set of detection probabilities (in Table 2), the left-hand panel of Figs 2 or 3 shows the plots of the average values (over 500 simulation trials) of the observed $Q_k$, the observed $Q_2$ and the estimated $\hat{Q}_2$ from solving eqn (7), as functions of the sample size; here, sample size refers to the number of sampling units, $T$. In the right-hand panel of each figure, we show the plots of the true species richness, the observed species richness, the Chao2 estimator based on the observed $Q_2$ [labelled Chao2 (observed $Q_2$) in the plots] and the derived species richness estimator based on the estimate $\hat{Q}_2$ [labelled Chao2 (estimated $Q_2$)].

Our simulation results reveal the following consistent patterns:

1. We first reiterate numerous prior findings (e.g. Colwell & Coddington 1994; Chao & Chiu 2012) regarding the performance of the Chao2 estimator based on the observed true counts of $Q_1$ and $Q_2$. As expected, the Chao2 estimator (the red dashed line in each of the right-hand panels) removes a large portion of the negative bias associated with the observed richness (the blue dash-dotted line), especially for small sample sizes or large CV values. As the number of sampling units is increased, the Chao2 estimator converges to the true species richness.

2. We next examine the performance of the species richness estimator calculated from the observed count of duplicates vs. the estimator calculated from the estimated count of duplicates based on solving eqn (7). From the left-hand panels of Figs 2 and 3, we can see that, for relatively small values of CV and higher mean detection probabilities (Case 1 to Case 5), our estimate of $Q_2$ is generally lower than the observed value of $Q_2$, leading to a higher Chao2 estimate [Chao2 (estimated $Q_2$)] than the estimate based on the observed $Q_2$ [Chao2 (observed $Q_2$)]. For the Random Parameters Group (Case 1 to Case 4, Fig. 2), the curve of the Chao2 estimator based on the estimated $Q_2$ is even closer to the true species richness line than the estimate based on the observed $Q_2$. By contrast, when the CV value is

### Table 2. The eight cases of detection probabilities $(\lambda_1, \lambda_2, \ldots, \lambda_{200})$ used in the simulations, in increasing order of true CV within two groups. Cases 1–4 apply random parameters and Cases 5–8 apply fixed parameters.

<table>
<thead>
<tr>
<th>Case</th>
<th>Detection probabilities $(\lambda_1, \lambda_2, \ldots, \lambda_{200})$</th>
<th>True CV</th>
<th>True mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>Random parameters group</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>$(\lambda_1, \lambda_2, \ldots, \lambda_{200})$-uniform (0, 0-2)</td>
<td>0.54</td>
<td>0.100</td>
</tr>
<tr>
<td>2</td>
<td>$(\lambda_1, \lambda_2, \ldots, \lambda_{200})$-negative binomial with size = 1, mean = 50, $\lambda_i = \frac{\alpha_i}{\max(\lambda_1, \lambda_2, \ldots, \lambda_{200})}$</td>
<td>0.88</td>
<td>0.169</td>
</tr>
<tr>
<td>3</td>
<td>$(\lambda_1, \lambda_2, \ldots, \lambda_{200})$-exponential with mean 1 (i.e. broken stick model), $\lambda_i = \frac{\alpha_i}{\max(\lambda_1, \lambda_2, \ldots, \lambda_{200})}$</td>
<td>1.13</td>
<td>0.103</td>
</tr>
<tr>
<td>4</td>
<td>$(\lambda_1, \lambda_2, \ldots, \lambda_{200})$-log normal with $\mu = \sigma^2 = 1, \lambda_i = \frac{\alpha_i}{\max(\lambda_1, \lambda_2, \ldots, \lambda_{200})}$</td>
<td>2.09</td>
<td>0.048</td>
</tr>
<tr>
<td>Fixed parameters group</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>$\lambda_1 = \lambda_2 = \ldots = \lambda_{200} = 0.2$</td>
<td>0</td>
<td>0.2</td>
</tr>
<tr>
<td>6</td>
<td>$\lambda_i = 1/i, i = 1, 2, \ldots, 200$ (Zipf-Mandelbrot model)</td>
<td>2.92</td>
<td>0.029</td>
</tr>
<tr>
<td>7</td>
<td>$\lambda_i = 1/i^{1.5}, i = 1, 2, \ldots, 200$ (Power-decay model)</td>
<td>3.54</td>
<td>0.023</td>
</tr>
<tr>
<td>8</td>
<td>$\lambda_i = 0.5/(i - 0.5), i = 1, 2, \ldots, 200$ (Zipf-Mandelbrot model)</td>
<td>4.22</td>
<td>0.018</td>
</tr>
</tbody>
</table>

True CV and mean; the CV and mean of $(\lambda_1, \lambda_2, \ldots, \lambda_{200})$. See Magurran (2004) for all models and distributions.

relatively large and mean detection probability is relatively low (Case 6 to Case 8, Fig. 3) our estimate of $Q_2$ is slightly higher than the observed value, leading to a lower Chao2 estimate than the estimate based on the observed $Q_2$, producing a larger bias, but with little difference once the number of sampling units becomes large.

Fig. 2. Simulation results for the four cases of detection probabilities in the Random Parameters Group (Cases 1–4) in Table 2. The left-hand panel (1a–4a) for each case shows the plots of the average values (over 500 simulation trials) of the observed $Q_1$ (blue dash-dotted line), the observed $Q_2$ (red dashed line) and the estimated $Q_2$ (black solid line) from solving eqn (7), as functions of the number of sampling units. The right-hand panel (1b–4b) of each figure shows the plots of the true species richness (horizontal black dotted lines), the observed species richness (blue dash-dotted lines), the Chao2 estimator based on the observed $Q_2$ (red dashed line) and the derived species richness estimator (black solid line) based on the estimated $Q_2$ from solving eqn (7).
Fig. 3. Simulation results for the four cases of detection probabilities in the Fixed Parameters Group (Cases 5–8) in Table 2. See the caption for Fig. 2 for explanation of the plots.
3. In summary, the curve of the Chao2 estimator using the estimated \( Q_2 \) deviates from that using the observed \( Q_2 \) to only a limited extent. As the number of sampling units is increased (with increased data collection), the discrepancy diminishes or vanishes. This result supports the assertion that our estimator (derived rigorously from Good–Turing theory) extracts useful information about the count of duplicates, from which we can legitimately infer species richness.

**Borneo fish data**

In the Borneo reef fish dataset (Fig. 1) described in the Introduction and Empirical Datasets sections, 441 species were observed over the course of \( T = 116 \) sampling sessions. The number of uniques (species seen in only one session) was \( Q_1 = 101 \), and the number of super-duplicates (species seen in two or more sessions) was \( Q_{2+} = 340 \). Substituting these data values with \( T = 116 \) into eqn (7) and solving for \( Q_2 \) using the R code supplied in Appendix S3 or the online software ‘SUPERDUP Licates’, we obtained an estimated number of duplicates \( \hat{Q}_2 = 49.9 \). Using this estimated number of duplicates and the observed number of uniques, we obtained a Chao2 estimate of 542.3 with an estimated standard error (SE) of 13.3 based on 200 bootstrap replications, yielding a 95% confidence interval for species richness of (519.4, 571.9) by means of a log transformation (Chao 1987). Our analysis implies that there were still undetected species despite the extensive survey, with the estimated number of the undetected species ranging from 78.4 to 130.9.

The above analysis uses \( T = 116 \) sampling units. We now examine, in Table 3, the sensitivity of our inferences to \( T \), by assuming various values of \( T \) in our analysis. When the sample size \( T \) was increased, the estimated count of duplicates generally decreased to a stable value around 49.8, and the corresponding Chao2 estimates tended to increase to 543.3 (SE = 13.8 based on 200 bootstrap replications) with a 95% confidence interval of (519.7, 574.1). These results differ little from the above results for \( T = 116 \). Table 3 reveals that, if \( T \) is large enough, the effect of \( T \) on our estimated count of duplicates is almost negligible, so that our species richness estimate is quite robust with respect to the value of \( T \). Consequently, as long as \( T \) is large, our estimation procedure can be applied even if \( T \) is unknown.

**Discussion and conclusions**

Until now, the minimum information required for the simplest species richness estimator, based on incidence data, required exact counts of uniques and duplicates and the number of observed species. We have developed a novel method, using the Good–Turing frequency formula, to estimate the true number of duplicates, and thus the true species richness, from the observed number of uniques and super-duplicates.

For abundance data, we have developed a parallel method that requires only the observed number of singletons and super-doubletons, to estimate the true richness of a single sample (Appendix S2). No higher frequency counts are needed, and even the total number of individuals in the sample need not be known precisely, as long as it is relatively large.

For incidence data, based on our tests on several real-world datasets with a known number of duplicates and on numerical studies on simulated data, we demonstrated that our proposed method works quite satisfactorily for a wide range of species detection probabilities. We can thus recommend it for practical use when an exact count of duplicates is impossible or requires substantial sampling effort, but uniques can be counted with reasonable confidence. For example, our method could be useful for inferring the species richness of communities for which the counting of duplicates is not feasible: a utility that we have demonstrated by analysing otherwise unusable data for coral reef fishes.

For abundance data, the single sample could be a pooling of informally collected data, including data collected by citizen scientists. Many people keep a ‘backyard list’ of birds or butterflies seen in a particular locality over a period of time. As long as the list indicates which species are singletons – a single individual seen on one occasion (a fact that an avid observer often knows), the true number of species in the long-term assemblage can be estimated with the tools detailed in Appendix S2. The total number of individuals seen over the course of the study need not be known, assuming it is reasonably large. For example, the 10-year bird list of citizen scientist C. P. Colwell (pers. comm.) covers 400 ha in Colorado, USA (2300–2500 m a.s.l., 39°22′N, 105°24′W). Colwell recorded 89 species, of which 25 were singletons. Abundances were not recorded. Because thousands of individuals were observed over the course of the study, however, our method can be applied, yielding an estimate of 115 species for the true richness of birds with a 95% confidence interval of (98.7, 157.5) in the study area.

**Table 3. The effect of sample size on the estimated count of duplicates, Chao2 estimates, standard error (SE) of the Chao2 estimate and 95% confidence intervals for the Borneo coral fish data (Townsend 2015)**

<table>
<thead>
<tr>
<th>Hypothetical value of ( T )</th>
<th>Estimate of ( Q_2 )</th>
<th>Chao2 estimate</th>
<th>Bootstrap SE of Chao2 estimate</th>
<th>95% bootstrap CI of species richness</th>
</tr>
</thead>
<tbody>
<tr>
<td>( T = 10 )</td>
<td>57.8</td>
<td>520.4</td>
<td>13.1</td>
<td>(498.6, 550.3)</td>
</tr>
<tr>
<td>( T = 20 )</td>
<td>52.4</td>
<td>533.4</td>
<td>14.0</td>
<td>(509.7, 565.1)</td>
</tr>
<tr>
<td>( T = 30 )</td>
<td>51.2</td>
<td>537.4</td>
<td>13.8</td>
<td>(513.9, 568.3)</td>
</tr>
<tr>
<td>( T = 50 )</td>
<td>50.3</td>
<td>540.3</td>
<td>13.2</td>
<td>(517.6, 569.7)</td>
</tr>
<tr>
<td>( T = 116 )</td>
<td>49.9</td>
<td>542.3</td>
<td>13.3</td>
<td>(519.4, 571.9)</td>
</tr>
<tr>
<td>( T = 200 )</td>
<td>49.8</td>
<td>542.8</td>
<td>14.1</td>
<td>(518.8, 574.4)</td>
</tr>
<tr>
<td>( T = 1000 )</td>
<td>49.8</td>
<td>543.3</td>
<td>12.7</td>
<td>(521.2, 571.3)</td>
</tr>
<tr>
<td>( T \to \infty )</td>
<td>49.8</td>
<td>543.3</td>
<td>13.8</td>
<td>(519.7, 574.1)</td>
</tr>
</tbody>
</table>

The results for \( T = 116 \) correspond to the actual data. Smaller values of \( T \) assess the sensitivity of the analysis to the number of sampling units. A bootstrap method (Appendix S2) was applied to obtain the SE of the Chao2 estimator and the associated CI via a log transformation. All results were obtained by means of the R code in Appendix S3 or the online software ‘SUPERDUP Licates’.
With growing acknowledgement and support from the academic community, public participation in scientific research is becoming increasingly prevalent, with many local, regional and global citizen science groups undertaking projects in various disciplines (e.g. ornithology, entomology, astronomy, environmental sciences and water quality, among many others). Our approach, based on simple incidence or abundance data, can be potentially applied to many types of citizen science data to provide statistically rigorous estimates of undetected species/class richness, using the online software we provide. The webpage of the software has been specifically designed (by and for) citizen scientists; see Appendix S5 for an introduction to the online software `SUPERDUPLICATES'. Given a list of species or classes observed, with only minimum additional effort to record the counts of singletons or unique, citizen scientists can easily estimate the number of previously undetected numbers of bird species in an area, moth species in a field, disease cases in plants/animals, stars in the sky or organic pollutants in an environmental unit. The information may further help scientists to efficiently allocate sampling effort and optimally determine sampling schemes. Because our estimators for duplicates (for incidence data) and doubletons (for abundance data) feed directly into the equations for Chao2 and Chao1 (Chao 1984), the rich theoretical and applied literature on these widely used richness estimators stands behind this new development (e.g. Hughes et al. 2001; Hill et al. 2003; Schloss & Handelsman 2005; Eren et al. 2016).

Authors’ contributions

A.C., R.K.C. and C.H.C. conceived the statistical ideas and designed the methodology; D.T. motivated the project and collected the principal data; all authors analysed the data and developed the software; A.C., R.K.C. and D.T. led the writing of the manuscript. All authors contributed critically to the drafts and gave final approval for publication.

Acknowledgements

The authors thank the Executive Editor (Robert Freckleton), an Associate Editor and two reviewers for providing helpful and thoughtful comments and suggestions. We are especially grateful for one reviewer's comment regarding a potential bias of our original approach to variance assessment and interval estimation, which led to the new bootstrap method in Appendix S2. This work is supported by the Taiwan Ministry of Science and Technology under Contract 103-2628-M-007-007 and 105-2628-M-007-001 (for A.C.) and 104-2118-M-007-006-MY3 (for C.H.C.). R.K.C. was supported by CAPES Ciência sem Fronteiras (Brazil). We thank Charles P. Colwell for sharing his bird census data.

Data accessibility

All data used in this manuscript are present in the manuscript and its supporting information.

References


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


Supporting Information

Details of electronic Supporting Information are provided below.

Appendix S1. Mathematical details.

Appendix S2. A bootstrap method to assess the variance of the proposed species richness estimator and to construct the associated confidence interval.

Appendix S3. R code for abundance/incidence data.

Appendix S4. Model and estimation for abundance data.

Appendix S5. An introduction to the application ‘SUPERDUPLICATES’.
Appendix

Seen once or more than once: applying Good-Turing theory to estimate species richness using only unique observations and a species list

Anne Chao, Robert K. Colwell, Chun-Huo Chiu, and Ditch Townsend

Appendix S1: Mathematical details

Proof of eqn 6 of the main text

Consider the incidence data for $T$ sampling units and the model proposed in eqn 1 of the main text, with species detection probabilities $\{\lambda_1, \lambda_2, \cdots, \lambda_S\}$. We first prove eqn 6 in the main text. Let $\lambda_{(1)}$ denote the mean detection probability of all unique species (species that are observed in only one sampling unit). Following typical Good-Turing theory, we assume that the detection probabilities of all unique species are approximately the same and equal to $\lambda_{(1)}$. In this case, other species could be highly heterogeneous without affecting the estimator. Chao et al. (2014, Appendix H) introduced the following estimator for $\lambda_{(1)}$, but did not provide a formal derivation,

$$\hat{\lambda}_{(1)} = \frac{2Q_2}{(T-1)Q_1 + 2Q_2}. \quad (A1)$$

We now present a proof of Eq. (A1). For simplicity, we define $I[A]$ as an indicator function, i.e., $I[A] = 1$ if $A$ is true, and $I[Y_{1}=1] = 0$. Using eqn 2 of the main text, we write

$$\frac{\hat{\lambda}_{(1)}}{1 - \hat{\lambda}_{(1)}} E(Q_1) \approx E \sum_{i=1}^{S} \frac{\lambda_i}{1 - \lambda_i} I[Y_i = 1]$$

$$= \sum_{i=1}^{S} \frac{\lambda_i}{1 - \lambda_i} T\lambda_i (1 - \lambda_i)^{-1} = \sum_{i=1}^{S} T\lambda_i^2 (1 - \lambda_i)^{-2}$$
\[
\frac{2}{T-1} \left[ \sum_{i=1}^{S} \left( \frac{T}{2} \right) \lambda_i^2 (1 - \lambda_i)^{T-2} \right] = \frac{2}{T-1} E(Q_2).
\]

Replacing the expected values by the observed data in the first and the last terms in the above, we then have \( \frac{\lambda_{(1)}}{1 - \lambda_{(1)}} Q_1 \approx \frac{2Q_2}{(T-1)} \), which leads to Eq. (A1).

**Proof of eqn 4 and eqn 7 of the main text**

Next, we provide proof of the following approximation, i.e., eqn 4 of the main text:

\[
\frac{E(Q_k)}{T \choose k} \approx (k-1)! \left( \frac{\lambda_{(1)}}{1 - \lambda_{(1)}} \right)^{k-1} \frac{E(Q_k)}{T}, \quad k = 2, \ldots, T.
\]

We first show that the above approximation formula is valid for the special case of \( k = 2 \).

From eqn 2 of the main text, we have

\[
\frac{E(Q_2)}{T \choose 2} = \sum_{i=1}^{S} \lambda_i^2 (1 - \lambda_i)^{T-2} = \sum_{i=1}^{S} \frac{\lambda_i}{1 - \lambda_i} \lambda_i (1 - \lambda_i)^{T-1}.
\]

Because \( P(Y_i = 1) = T \lambda_i (1 - \lambda_i)^{T-1} \), the above summation becomes

\[
\sum_{i=1}^{S} \frac{\lambda_i}{1 - \lambda_i} \lambda_i (1 - \lambda_i)^{T-1} = \sum_{i=1}^{S} \frac{\lambda_i}{1 - \lambda_i} \frac{1}{T} [T \lambda_i (1 - \lambda_i)^{T-1}]
\]

\[
= \frac{1}{T} \sum_{i=1}^{S} \frac{\lambda_i}{1 - \lambda_i} P(Y_i = 1) = \frac{1}{T} E \left\{ \sum_{i=1}^{S} \frac{\lambda_i}{1 - \lambda_i} I[Y_i = 1] \right\}.
\]

Here \( I[Y_i = 1] \) is an indicator function, as defined earlier. Further, note in the sum

\[
\sum_{i=1}^{S} \frac{\lambda_i}{1 - \lambda_i} J[Y_i = 1]
\]

that only unique species \( i \) would contribute the term \( \frac{\lambda_i}{1 - \lambda_i} \), whereas other species do not contribute anything. Under our assumption that all unique species have approximately the same probabilities (Turing’s assumption), concentrated at \( \lambda_{(1)} \), each unique
species would contribute \( \frac{\lambda_{(i)}}{1 - \lambda_{(i)}} \) and there are \( Q_1 \) of them. Consequently, we have

\[
\sum_{i=1}^{S} \frac{\lambda_i}{1 - \lambda_i} I[Y_i = 1] \approx Q_1 \times \frac{\lambda_{(1)}}{1 - \lambda_{(1)}},
\]

implying

\[
\sum_{i=1}^{S} \frac{\lambda_i}{1 - \lambda_i} \lambda_i(1 - \lambda_i)^{-1} \approx \frac{1}{T} E\left[Q_i \times \frac{\lambda_{(1)}}{1 - \lambda_{(1)}}\right] = \frac{\lambda_{(1)}}{1 - \lambda_{(1)}} E\left(\frac{Q_1}{T}\right).
\]

This leads to the following approximation:

\[
E\left(\frac{Q_2}{T}\right) \approx \left(\frac{\lambda_{(1)}}{1 - \lambda_{(1)}}\right) \frac{E\left(\frac{Q_1}{T}\right)}{T}.
\]

For the case of \( k = 3 \), parallel derivations yield:

\[
E\left(\frac{Q_3}{T}\right) \approx \left(\frac{\lambda_{(1)}}{1 - \lambda_{(1)}}\right) \left(\frac{\lambda_{(2)}}{1 - \lambda_{(2)}}\right) \frac{E\left(\frac{Q_1}{T}\right)}{T}.
\]

Similarly, for \( k = 4 \), we have

\[
E\left(\frac{Q_4}{T}\right) \approx \left(\frac{\lambda_{(1)}}{1 - \lambda_{(1)}}\right) \left(\frac{\lambda_{(2)}}{1 - \lambda_{(2)}}\right) \left(\frac{\lambda_{(3)}}{1 - \lambda_{(3)}}\right) \frac{E\left(\frac{Q_1}{T}\right)}{T}.
\]

The general approximation formula turns out to be:

\[
E\left(\frac{Q_k}{T}\right) \approx \prod_{j=2}^{k} \left(\frac{\lambda_{(j)}}{1 - \lambda_{(j)}}\right) \frac{E\left(\frac{Q_1}{T}\right)}{T}, \quad k = 2, \ldots, T.
\]

Notice that the mean detection probability of species that appeared in \( k \) sampling units is approximately \( k \) times of that of the unique species, i.e.,

\[
\frac{\lambda_{(k)}}{1 - \lambda_{(k)}} \approx k \left(\frac{\lambda_{(1)}}{1 - \lambda_{(1)}}\right), \quad k = 2, \ldots, T.
\]
We then obtain the following approximation (eqn 4 in the main text):

\[
E(Q_k) \approx \binom{T}{k}^{-1} \frac{E(Q_1)}{T} (k-1)! \left( \frac{\lambda_{(1)}}{1-\lambda_{(1)}} \right)^{k-1}, \quad k = 2, \ldots, T,
\]

which subsequently leads to

\[
E(Q_{z^+}) = \frac{E(Q_1)}{T} \left[ \frac{T}{2} \left( \frac{\lambda_{(1)}}{1-\lambda_{(1)}} \right) + 2! \left( \frac{T}{3} \left( \frac{\lambda_{(1)}}{1-\lambda_{(1)}} \right)^2 \right) + \cdots + (T-1)! \left( \frac{T}{T-1} \left( \frac{\lambda_{(1)}}{1-\lambda_{(1)}} \right)^{T-1} \right) \right].
\]

Substituting the formula \( \lambda_{(1)} \) in Eq. (A1) into the above and replacing the expected values by the observed data, we obtain

\[
Q_{z^+} \approx Q \left( \frac{2Q_2}{Q_1} \right)^1 \left( \frac{2Q_2}{Q_1} \right)^2 \cdots \left( \frac{2Q_2}{Q_1} \right)^T.
\]

When \( T \) is not large, the R code supplied in Appendix S3 can be applied to obtain all numerical results for any given \( T, Q_1 \) and \( Q_{z^+} \). If observed \( T \) is large enough, we can let \( T \) tend to infinity (the tail sequences have a negligible sum) and the above equation can be simplified to eqn 7 in the main text:

\[
Q_{z^+} \approx \frac{Q_1^2}{2Q_2} \left[ -\log \left( 1 - \frac{2Q_2}{Q_1} \right) - \frac{2Q_2}{Q_1} \right].
\]

As discussed in the main text, the solution is unique and can be obtained using the R code (Appendix S3) or the online software “SuperDuplicates” available from https://chao.shinyapps.io/SuperDuplicates/.

Reference


Appendix

Seen once or more than once: applying Good-Turing theory to estimate species richness using only unique observations and a species list

Anne Chao, Robert K. Colwell, Chun-Huo Chiu, and Ditch Townsend

Appendix S2: A bootstrap method to assess the variance of the proposed species richness estimator and to construct the associated confidence interval

We propose the use of a bootstrap method to approximate the variance of the proposed richness estimator and to construct the associated confidence interval. Here we present the bootstrap procedures for incidence data; parallel steps can be formulated for abundance data. Our proposed bootstrap procedures are described as follows.

1. Given the number of uniques ($Q_1$) and the number super-duplicates ($Q_{2\tau} = S_{obs} - Q_1$) in the observed data, we first obtain the estimated count of duplicates from eqn 7 of the main text. Then, based on the estimated count of duplicates, along with $Q_1$, we obtain the proposed species richness estimator ($\hat{S}_{Chao}$) from eqn 3 of the main text.

2. Consider a “bootstrap assemblage” that consists of $\hat{S}_{Chao2}$ species; each species is classified into one of the three categories (unobserved, uniques, super duplicates) with cell probabilities

$$\left\{ 1 - \frac{S_{obs}}{\hat{S}_{Chao2}}, \frac{Q_1}{\hat{S}_{Chao2}}, \frac{S_{obs} - Q_1}{\hat{S}_{Chao2}} \right\}.$$ 

3. Generate a multinomial random vector $(Q_0^*, Q_1^*, Q_{2\tau}^*)$ from the bootstrap assemblage, with cell total $\hat{S}_{Chao2}$ and cell probabilities

$$\left\{ 1 - \frac{S_{obs}}{\hat{S}_{Chao2}}, \frac{Q_1}{\hat{S}_{Chao2}}, \frac{S_{obs} - Q_1}{\hat{S}_{Chao2}} \right\}.$$ 

Then based
on the bootstrap data \((Q_1^*, Q_2^*)\), which mimic the observed data in a bootstrap replication, we can obtain a new count of duplicates and thus a new bootstrap species richness estimator. (The uncertainty due to the estimation of the count of duplicates can be incorporated in this procedure.)

(4) Replicate Step (3) \(B\) times, and obtain \(B\) bootstrap species richness estimates \((B = 200\) or larger is suggested). The bootstrap variance estimator of the proposed species richness estimator is the sample variance of these \(B\) values. The resulting bootstrap variance is then used to construct a bootstrap confidence interval via a log-transformation (Chao 1987).

References

Appendix

Seen once or more than once: applying Good-Turing theory to estimate species richness using only unique observations and a species list

Anne Chao, Robert K. Colwell, Chun-Huo Chiu, and Ditch Townsend

Appendix S3: R code for abundance/incidence data

SuperDuplicate=function (Q1,Sobs,T=0)
{
  #input:
  #Q1 is the number of unique species for incidence data (or singletons for abundance data).
  #Sobs is the observed species richness.
  #T is the known number of sampling units for incidence data; for incidence data with
  #unknown T or for abundance data, set T = 0 (default).

  ####### Estimation of the number of duplicates or doubletons
  Q2.est=function (Q1,Sobs,T=0)
  {
    Q2_=Sobs-Q1;
    if(T<1){
      F1=function(f2,f2_,f1) f1^2/(2*f2)*(log(f1/(f1-2*f2))-2*f2/f1)-f2_;
      t1=0.0001;t2=Q1/2;t=(t1+t2)/2;
      while(abs(t-t1)>0.05 || abs(t-t2)>0.05){
        K1=F1(t,Q2_,Q1);K2=F1(t1,Q2_,Q1);K3=F1(t2,Q2_,Q1);
        if(K1*K2<0) {t1=t1;t2=t;t=(t+t1)/2;}
    }
if(K1*K3<0) {t1=t;t2=t2;t=(t+t2)/2;}
}
}else{

F1=function(f2,f2_,f1,T) f1^2/(2*f2)*sum((2*f2/f1)^2:(2:T))/(2:T)-f2_

\[
t1=0.0001; t2=Q1; t=(t1+t2)/2;
\]

while(abs(t-t1)>0.05 || abs(t-t2)>0.05){

K1=F1(t,Q2_,Q1,T); K2=F1(t1,Q2_,Q1,T); K3=F1(t2,Q2_,Q1,T);

if(K1*K2<0) {t1=t1; t2=t; t=(t+t1)/2;}

if(K1*K3<0) {t1=t; t2=t2; t=(t+t2)/2;}
}
}

return(t)


######## Bootstrap variance of species richness estimator

V,Q2=function(Q1,Sobs,T=0){

Q1=ifelse(Q1==0,1,Q1);

A=ifelse(T<1,1,(T-1)/T);

Q2=ifelse(Q1==Sobs,0,Q2.est(Q1,Sobs,T))

if(round(Q2)>1){ chao22=Sobs+A*Q1^2/(2*Q2);
}
else{ chao22=Sobs+A*(Q1-1)*Q1/2 }

chao2=round(chao22);

times=200; eS=numeric(times); pS=Sobs/chao2; p1=Q1/chao2;

for(j in 1:times){

data=rmultinom(1, size=chao2, prob=c(pS-p1, p1, 1-pS))

temp2=ifelse(tempS>temp1,Q2.est(temp1,tempS,T),0);
if(round(temp2)>1){ eS[j]=tempS+A*temp1^2/(2*temp2);
}else{eS[j]=tempS+A*(temp1-1)*temp1/2 
}
return(var(eS))

########## Main code
Q2_=Sobs-Q1;
Q2=Q2.est(ifelse(Q1==0,1,Q1),Sobs,T);

A=ifelse(T<1,1,(T-1)/T);
if(round(Q2)>0){ Chao2=Sobs+A*Q1^2/(2*Q2);
}else{Chao2=Sobs+A*(Q1-1)*Q1/2}

V=V.Q2(Q1,Sobs,T)
R=exp(1.96*(log(1+V/(Chao2-Sobs)^2))^(1/2))
L=Sobs+(Chao2-Sobs)/R;
U=Sobs+(Chao2-Sobs)*R;
output=matrix(0,nrow=1,ncol=7)
output[1,]=c(Sobs,Q1,Q2,Chao2,V^(1/2),U,L);
colnames(output)=c("Sobs","Q1","Q2.est","Chao2.est","sd","95% C.I. Upper","95% C.I. Lower")
return(output)
Appendix

Seen once or more than once: applying Good-Turing theory to estimate 
species richness using only unique observations and a species list

Anne Chao, Robert K. Colwell, Chun-Huo Chiu, and Ditch Townsend

Appendix S4: Model and estimation for abundance data

Data and Model
Consider a community consisting of $N$ total individuals, each belonging to one of $S$ distinct 
species. Let $N_i$ be the number of individuals of the $i$th species (true absolute abundance), $i = 1, 2, \cdots, S$, $N_i > 0$, and $N = \sum_{i=1}^{S} N_i$. The relative abundance $p_i$ of species $i$ is $p_i = N_i / N$, so that 
$\sum_{i=1}^{S} p_i = 1$. Here $N$, $S$, $N_i$, and $p_i$ represent the true, unknown parameters in the community.

For abundance data, we assume that a random sample of $n$ individuals (sample size $n$ is 
fixed in advance) is taken from the community and a total of $S_{\text{obs}}$ species is observed in the 
sample. Let $X_i$ (sample species frequency) be the number of individuals of the $i$th species that 
are observed in the sample, $i = 1, 2, \cdots, S$. Only those species with $X_i > 0$ are observable in the 
sample, and $\sum_{i=1}^{S} X_i = n$ (only species with $X_i > 0$ contribute to the sum).

Generally, species detection probability or detection rate for any sampled individual is a 
combination of species abundance $p_i$ and individual detectability. We assume that the 
detectability of any individual of the $i$th species is $\theta_i > 0$, which varies among species, and is 
determined by many possible factors (such as movement patterns, color, size, and 
vocalizations). Under this general model, the species detection probability for the $i$th species 
in any observation becomes $\psi_i = p_i \theta_i / \sum_{k=1}^{S} p_k \theta_k$, $i = 1, 2, \cdots, S$. Our model is the following 
multinomial model that allows heterogeneous species abundance and heterogeneous
individual detectability among species:

\[ P(X_1 = x_1, \ldots, X_S = x_S) = \frac{n!}{x_1! \cdots x_S!} \psi_1^{x_1} \psi_2^{x_2} \cdots \psi_S^{x_S}. \]

Here, the detection probability \( \psi_i \) is a normalized product of species relative abundance \( p_i \) and individual detectability \( \theta_i \). In the special case that individuals of all species have the same detectability, then \( \psi_i \) reduces to \( p_i \); the species detection probability for the \( i \)th species is simply the true relative abundance \( p_i = N_i / N \).

From the above general multinomial distribution, the sample frequency \( X_i \) of species \( i \) is distributed as a binomial random variable characterized by \((n, \psi_i)\), i.e.,

\[ P(X_i = x_i) = \binom{n}{x_i} \psi_i^{x_i} (1 - \psi_i)^{n-x_i}. \]  

(D1)

Comparing this model with the corresponding model for incidence data (eqn 1 in the main text), we see that the two types of data are based on the same structure of distributions.

Let \( f_k \) (abundance frequency counts), \( k = 0, 1, \ldots, n \), be the number of species represented by exactly \( k \) individuals in the sample. Thus, we have \( n = \sum_{i=1}^{S} X_i = \sum_{k \geq 1} k f_k \), and

\[ S_{obs} = \sum_{k \geq 1} f_k \]. In particular, \( f_1 \) is the number of species represented by exactly one individual (singletons) in the sample, \( f_2 \) is the number of species represented by exactly two individuals (doubletons), and \( f_0 \) denotes the number of undetected species in the sample. Here “undetected species” means species that are present in the community of \( N \) individuals and \( S \) species, but were not detected in the observed sample of \( n \) individuals and \( S_{obs} \) species. Because \( S = S_{obs} + f_0 \), species richness estimation is equivalent to the inference about the number of undetected species \( f_0 \).
Statistical Estimation

The statistical inference problem involved in this study is to estimate species richness based on the information on the number of singletons $f_1$ and the number super-doubletons $f_{2+} = f_2 + f_3 + \ldots + f_n$ (the number of species observed at least twice in the sample). For abundance data, the sample size $n$ is often very large. As with incidence data, our inference is very robust to the value of $n$. Based on the binomial model in Eq. (D1), we have a general form for the expected value of $f_k$:

$$E(f_k) = E\left(\sum_{i=1}^{S} I(X_i = k)\right) = \sum_{i=1}^{S} \binom{n}{k} \psi_i^k (1 - \psi_i)^{n-k}, \quad k = 0, 1, \ldots, n. \quad (D2)$$

Chao (1984) derived a lower bound of undetected species richness in terms of the numbers of singletons and doubletons; the corresponding lower bound of species richness given below is referred to as the Chao1 estimator: (Colwell & Coddington 1994)

$$\hat{S}_{\text{Chao}} = \begin{cases} 
S_{\text{obs}} + [(n-1)/n]f_1^2 / (2f_2), & \text{if } f_2 > 0, \\
S_{\text{obs}} + [(n-1)/n]f_1(f_1-1)/2, & \text{if } f_2 = 0. 
\end{cases} \quad (D3)$$

Based only on information on $f_1$ and $f_{2+}$, we need to first estimate $f_2$ to apply the above Chao1 richness estimator. Let $\psi_{(1)}$ denote the mean detection probabilities of all singletons.

Following parallel derivations as those in Appendix S1 for incidence data, we can express each of $E(f_2), E(f_3), \ldots, E(f_n)$ in terms of $\psi_{(1)}$ and $E(f_1)$ as follows:

$$\frac{E(f_k)}{\binom{n}{k}} \approx (k-1)! \left( \frac{\psi_{(1)}}{1-\psi_{(1)}} \right)^{k-1} \frac{E(f_1)}{n}, \quad k = 2, \ldots, n. \quad (D4)$$

Therefore, we have

$$E(f_{2+}) = \frac{E(f_1)}{n} \left[ \binom{n}{2} \left( \frac{\psi_{(1)}}{1-\psi_{(1)}} \right) + 2! \binom{n}{3} \left( \frac{\psi_{(1)}}{1-\psi_{(1)}} \right)^2 + \cdots + (n-1)! \binom{n}{n} \left( \frac{\psi_{(1)}}{1-\psi_{(1)}} \right)^{n-1} \right]. \quad (D5)$$
The Good-Turing frequency formula (Good 1953, 2000) states that $\psi_{(1)}$ is not estimated by its sample frequency $1/n$, but rather by $r/n$, where $r = 2f_2 / f_1$. An improved estimator takes the following form (e.g., see Chao & Jost 2012):

$$\hat{\psi}_{(1)} = \frac{2f_2}{(n-1)f_1 + 2f_2}.$$  \hspace{1cm} (D6)

Substituting Eq. (D6) into Eq. (D5) and replacing the expected values by the observed data, we obtain

$$f_{2+} \approx f_1 \left( \frac{f_1}{2f_2} \left[ \frac{1}{2} \left( \frac{2f_2}{f_1} \right)^2 + \frac{1}{3} \left( \frac{2f_2}{f_1} \right)^3 + \cdots + \frac{1}{n} \left( \frac{2f_2}{f_1} \right)^n \right] \right).$$  \hspace{1cm} (D7)

For abundance data, the sample size is usually large, so we can let $n$ tend to infinity in Eq. (D7) and obtain the following approximation:

$$f_{2+} \approx h(f_2) \equiv \frac{f_2^2}{2f_2} \left[ -\log \left( 1 - \frac{2f_2}{f_1} \right) - \frac{2f_2}{f_1} \right].$$  \hspace{1cm} (D8)

As with the incidence data, the solution is unique given $f_1$ and $f_{2+}$ because the function $h(f_2)$ is a strictly increasing function from 0 to infinity, when $f_2$ is increased from 0 to any large integer. The solution of Eq. (D8) and the resulting species richness estimator along with bootstrap s.e. and confidence intervals can be obtained using the R code (Appendix S3) or the online software “SuperDuplicates” available from https://chao.shinyapps.io/SuperDuplicates/.

**Numerical Test**

To test our proposed estimator for the doubleton count (Eq. B8), we selected eight abundance datasets with known doubleton counts. These eight datasets include the species abundances for samples of tree seedlings (Butler & Chazdon 1998), butterflies (Fisher et al. 1943), vascular plants (Miller & Wiegert 1989), trees (Hubbell et al. 2005), moths (Lewis & Taylor 1967), fishes (Dahlberg & Odum 1970), human phenotypic profiles and animal phenotypic
profiles (Mather et al. 2012) in increasing order of CV estimates; see Table D1 for details. For these datasets, the estimated CV values lie between 1.49 and 6.93, signifying moderate to high degrees of heterogeneity among species. As with incidence data, the CV estimator usually exhibits negative bias, the true CV in each case is most likely to be higher than the estimated CV.

Our findings for abundance data are generally consistent with those discussed in the main text for incidence data. That is, the Chao1 richness estimates computed based on the estimated doubleton counts generally yielded results quite close to the corresponding richness estimates based on the true doubleton counts. Except for the last three datasets, for which the CV value is the largest among the five assemblages, the proposed estimated doubleton count is somewhat lower than the true value, leading to somewhat higher Chao1 estimates than that based on the true doubleton count. However, the differences are limited, revealing that the proposed estimation generally works satisfactorily.
**Table D1.** Numerical test results of the proposed estimator for the count of doubletons for eight datasets in an increasing order of the estimated CV value. $n$: number of individuals; $S_{obs}$: number of observed species richness; CV estimate: the estimated CV (coefficient of variation) of detection probabilities among species; observed $f_1$: the observed true count of singletons; observed $f_2$: the observed true count of doubletons; estimated $f_2$: the estimated count of doubletons from solving Eq. (D8). 95% C. I. (confidence interval) for Chao1 with observed $f_2$: based on the Chao (1987) asymptotic variance formula via a log-transformation; 95% C. I. for Chao1 with estimated $f_2$: based on a bootstrap method with 200 replications (Appendix S2) via a log-transformation. All results were obtained by means of the R code in Appendix S3 or the online software “SuperDuplicates”.

<table>
<thead>
<tr>
<th>Data</th>
<th>$n$</th>
<th>$S_{obs}$</th>
<th>CV estimate</th>
<th>Observed $f_1$</th>
<th>Observed $f_2$</th>
<th>Estimated $f_2$</th>
<th>Chao1 with observed $f_2$ (95% C. I.)</th>
<th>Chao1 with estimated $f_2$ (95% bootstrap C. I.)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tree seedlings</td>
<td>952</td>
<td>34</td>
<td>1.49</td>
<td>2</td>
<td>2</td>
<td>1.0</td>
<td>35.0 (34.1, 45.1)</td>
<td>36.0 (34.4, 43.7)</td>
</tr>
<tr>
<td>Butterflies</td>
<td>9031</td>
<td>620</td>
<td>1.56</td>
<td>118</td>
<td>74</td>
<td>58.7</td>
<td>714.1 (681.1, 765.0)</td>
<td>738.6 (714.8, 768.5)</td>
</tr>
<tr>
<td>Plants</td>
<td>1008</td>
<td>188</td>
<td>1.78</td>
<td>61</td>
<td>35</td>
<td>28.8</td>
<td>241.2 (217.6, 283.6)</td>
<td>252.4 (231.6, 283.4)</td>
</tr>
<tr>
<td>Trees</td>
<td>242118</td>
<td>320</td>
<td>4.13</td>
<td>17</td>
<td>13</td>
<td>13.5</td>
<td>331.1 (323.7, 353.4)</td>
<td>347.0 (336.0, 365.6)</td>
</tr>
<tr>
<td>Moths</td>
<td>6815</td>
<td>197</td>
<td>4.38</td>
<td>37</td>
<td>22</td>
<td>18.4</td>
<td>228.1 (211.7, 262.7)</td>
<td>234.2 (220.4, 256.0)</td>
</tr>
<tr>
<td>Fishes</td>
<td>31637</td>
<td>70</td>
<td>4.69</td>
<td>14</td>
<td>5</td>
<td>6.95</td>
<td>89.6 (77.6, 128.3)</td>
<td>84.1 (76.4, 101.1)</td>
</tr>
<tr>
<td>Phynotypic profiles (Human)</td>
<td>2761</td>
<td>52</td>
<td>6.54</td>
<td>26</td>
<td>10</td>
<td>10.3</td>
<td>85.8 (66.7, 129.8)</td>
<td>84.6 (68.4, 116.8)</td>
</tr>
<tr>
<td>(Animal)</td>
<td>2439</td>
<td>35</td>
<td>6.93</td>
<td>16</td>
<td>5</td>
<td>6.7</td>
<td>60.6 (44.3, 105.7)</td>
<td>54.0 (43.8, 76.0)</td>
</tr>
</tbody>
</table>
References


Appendix

Seen once or more than once: applying Good-Turing theory to estimate species richness using only unique observations and a species list

Anne Chao, Robert K. Colwell, Chun-Huo Chiu, and Ditch Townsend

Appendix S5: An introduction to Application SuperDuplicates

Application SuperDuplicates (https://chao.shinyapps.io/SuperDuplicates/)

Anne Chao, Robert K. Colwell, Chun-Huo Chiu, and Ditch Townsend

If you know how many species have been seen in a given area, and if you know how many of those have only been seen once, this easy little tool will tell you approximately how many species are left to find. Amazingly, you don't have to know how many more times you have seen each species. Don't be misled by the apparent simplicity of this tool. It is based on a rigorous extension of mathematical work originally developed by Alan Turing and I. J. Good in their code-breaking work during WWII. Full details can be found in the following scientific paper:


What kind of data do you have?

Incidence data record the presence or absence of each observed species in repeated samples, called sampling units. The sampling units can be observations on different days or in different years, or from different dives, point-counts, traps, plots, or quadrats. A species observed in only one sampling unit is called a unique species. It doesn't matter how many individuals of that species were observed in that one sampling unit; the species is still a unique. A species observed in exactly two sampling units (regardless of the number of individuals) is called a duplicate. A super-duplicate is any species observed in more than one sampling units (but you don't have to know exactly how many more).

Abundance data record the number of individuals of each observed species in a single sample, or in a combined sample from several sampling units. A species with only a single individual in the sample is called a singleton. A species with exactly two individuals is a doubleton. A super-doubleton is any species with more than one individual in the sample (but again, you don't have to know exactly how many more).

What this tool does
This tool allows you to estimate the true number of species in a locality, based on a simple list of species observed, when you also know the number of uniques or singletons on the list. You do not need to know the number of duplicates, doubletons, or any other counts. This approach is based on the Turing's wisdom that the number of uniques/singletons contains nearly all the information about the number of undetected species.

**How to use this tool:**

**Step 1.** Select data type (incidence data or abundance data; see above).

**Step 2.**

(1) If you select Incidence Data, you need to enter the number of species in your list (Sobs), the number of unique species in the list (Q1, i.e., the number species that were each found in a single sampling unit), and the number of sampling units (if this number is unknown, enter 0 instead).

(2) If you select Abundance Data, you need to enter the total number of species in your list (Sobs) and the number of singletons in the list (f1, i.e., the number of species represented by exactly one individual).

**Step 3.** Press the <Run!> button. It may take several minutes for the results to appear, because estimating the confidence interval requires a complicated resampling procedure.

**Example:**

For the Borneo reef fish dataset discussed in Chao et al. (2017) paper, select data type and enter data as shown in the following input window. Then press the <Run!> button.
The following output will be shown under the “output” menu. (Note the bootstrap standard error and 95% confidence intervals will vary very slightly each time you enter the same data due to random bootstrap process.)

<table>
<thead>
<tr>
<th>Q2.est</th>
<th>Chao2.est</th>
<th>se</th>
<th>95% C.I. lower</th>
<th>95% C.I. upper</th>
<th>Undetected # species</th>
<th>Undetected percentage (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>49.86</td>
<td>542.42</td>
<td>14.40</td>
<td>517.89</td>
<td>574.76</td>
<td>101.42</td>
<td>18.70</td>
</tr>
</tbody>
</table>

- **Q2.est**: The estimated number of duplicates (species recorded in exactly two sampling units or sessions). This estimate is needed to obtain the species richness estimate.
- **Chao2.est**: Estimated number of species (species richness) for incidence data, including species present but not detected (Chao 1987).
- **se**: Standard error (sampling uncertainty) of the Chao2 estimator. This number is used to compute a plausible range (95% confidence interval) of species richness.
- **95% C.I. lower**: Lower bound of the estimated 95% confidence interval (C.I.) of species richness.
- **95% C.I. upper**: Upper bound of the estimated 95% confidence interval (C.I.) of species richness.
- **Undetected # species**: The estimated number of species present but not yet seen.
- **Undetected percentage**: The percentage of species present but not (yet) seen.

**NOTE**: Because this tool runs a random bootstrapping process in the background, with 200 loops, to estimate standard error, the output (for se, 95% C. I. lower, and 95% C. I. upper) will vary very slightly each time you enter the same data.