Richness estimation with species identity error

Jai-Hua Yen* and Chun-Huo Chiu
Department of Agronomy, National Taiwan University, Taipei 10617, Taiwan.
*r06621209@ntu.edu.tw, chchiu2017@ntu.edu.tw

National Taiwan University

Introduction

- Completely species inventories in the wild field are almost unattainable goals. Hence, the observed richness in the sample always underestimates the true species richness in the assemblage.
- Species identity error almost occurred in every survey especially in vegetation sampling was recently discussed in the literatures (Burg *et al*., 2015; Morrison, 2015).
- Without error correction, the richness estimation will be inaccurate based on original sampling data.

Conclusions

- When species identity error occurs, the estimation of richness estimator would seriously underestimate the true richness even though the increase of sampling units.
- We suggest that the adjusted richness estimator should be applied to estimate species richness of the target region.

Methodology

Notation

- e: species identity error rate
- r: mean probability that a species is misidentified into another species belongs to the sampling plot
- T: number of sampling units
- $S_{obs,e}$: number of species recorded in the sample when species identity error occurs
- Q_{ke}: number of species recorded exactly k times in the sample when species identity error occurs

Step 1: Estimate mean species identity error rate

- Number of species (S_{sub}) and categories of species in the subplot must known by the experiment designer, but unknown by the observer.
- The expectation of the number of observed species does not belong to the subplot ($f_{sub,0}$) is:

$$E\left(f_{sub,0}\right) \approx S_{sub} \times \bar{e} \times (1-r).$$

- The expectation of the number of observed species belongs to the subplot ($S_{sub,e}$) is:

$$E\left(S_{sub,e}\right) \approx S_{sub} - S_{sub} \times \bar{e} \times \left(1 - \frac{\bar{e}}{\frac{S_{sub}}{r} - 1}\right)^{S_{sub}-1}.$$

- By solving those two equations, we have the estimate of $\bar{e}$ and r which are denoted by $\hat{\bar{e}}$ and $\hat{r}$.

Step 2: Adjustment species richness estimator

- When species identity error occurs, the observed, singleton, and doubleton richness adjustments are

$$S_{obs,a} = \frac{S_{obs,e}}{1 - \hat{\bar{e}} \times \hat{r}}.$$

$$Q_{1a} = \frac{Q_{1e}}{(1 - \hat{\bar{e}} \times \hat{r}) \times exp(-\hat{\bar{e}} \times \hat{r})}.$$

$$Q_{2a} = \frac{Q_{2e} - Q_{1a} \times \hat{\bar{e}} \times \hat{r} \times \left(1 - \frac{1}{T}\right) \times \frac{Q_{1a}}{S_{obs,a}} \times exp(-\hat{\bar{e}} \times \hat{r})}{(1 - \hat{\bar{e}} \times \hat{r}) \times exp(-\hat{\bar{e}} \times \hat{r})}.$$

- The corrected Chao2 estimator is

$$\hat{S}_{Chao2,c} = S_{obs,a} + \frac{T-1}{T} \frac{Q_{1a}^2}{2Q_{2a}}.$$

- When $0 \le Q_{2a} < 1$, corrected richness estimator is

$$\hat{S}_{Chao2,c} = S_{obs,a} + \frac{T-1}{T} \frac{Q_{1a}(Q_{1a}-1)}{2(Q_{2a}+1)}.$$

- An adjusted richness estimator is proposed to deal with the inaccuracy of corrected Chao2 estimator

$$\hat{S}_{adj} = S_{obs,a} + \frac{T-1}{T} max\left\{\left(\frac{Q_{1a}^2}{2Q_{2a}} - \frac{Q_{1a}}{2Q_{2a}} - \frac{Q_{1a}^2}{2Q_{2a}^2}\right), 0\right\}.$$

- When $0 \le Q_{2a} \le 1$, the adjusted richness estimator applied first-order Jackknife estimator (Burnham & Overton, 1978)

$$\hat{S}_{adj} = S_{obs,a} + \frac{T-1}{T} Q_{1a}.$$

Results

Simulation Studies

- True method: use Chao2 estimator by the data without species identity error.
- Observed method: use Chao2 estimator by the data with species identity error.
- Adjusted method: use adjusted richness estimator by the data with species identity error.

Table 1.
Comparison of species richness estimator for incidence data based on 500 simulation data sets and 200 bootstrapping trials under random uniform (0, 1) model, with $\bar{p} = 0.51$, $CV = 0.53$, $S = 100$, $S_{sub} = 40$, $T = 5$, and $r = 0.91$.

Mean error rate	Estimated error rate	Method	S_{obs}	Q_1	Q_2	$\hat{S}$	Bias	Sample s.e.	Estimated s.e.	Sample RMSE
0	0	True	85.2	15.3	17.3	91.37	-8.63	4.82	4.19	9.89
0.053	0.058	Observed	81.5	13.9	15.8	87.22	-12.78	5.46	4.06	13.9
		Adjusted	86.3	15.6	17.5	92.05	-7.95*	7.17	8.33	10.71†
0.097	0.098	Observed	78.3	13.2	14.8	83.72	-16.28	5.29	3.95	17.12
		Adjusted	86.3	15.9	17.5	92.2	-7.8*	7.92	9.4	11.12†
0.15	0.157	Observed	74	11.7	13.4	78.86	-21.14	5.24	3.75	21.78
		Adjusted	86.8	16	17.6	92.89	-7.11*	10.33	10.2	12.54†
0.199	0.209	Observed	70.7	10.3	12.7	74.71	-25.29	5.01	3.34	25.78
		Adjusted	88.3	15.8	18.5	94.34	-5.66*	14.05	11.12	15.15†

*Denotes the smaller bias. †Denotes the smaller RMSE.

Table 2.
Comparison of species richness estimator for incidence data based on 500 simulation data sets and 200 bootstrapping trials under random uniform (0, 1) model, with $\bar{p} = 0.51$, $CV = 0.53$, $S = 100$, $S_{sub} = 40$, $T = 20$, and $r = 0.91$.

Mean error rate	Estimated error rate	Method	S_{obs}	Q_1	Q_2	$\hat{S}$	Bias	Sample s.e.	Estimated s.e.	Sample RMSE
0	0	True	95.3	4.1	3.9	98.8	-1.2	4.9	4.25	5.06
0.053	0.055	Observed	91.2	3.9	3.6	94.8	-5.2	5.46	4.45	7.53
		Adjusted	96.1	4.3	4	97.85	-2.15*	5.26	5.39	5.68†
0.097	0.095	Observed	87.3	3.3	3.5	90.1	-9.9	5.15	3.76	11.15
		Adjusted	95.8	4	4.1	97.1	-2.9*	6.52	5.72	7.14†
0.15	0.151	Observed	82.9	3.1	2.9	85.61	-14.39	5.21	3.79	15.31
		Adjusted	96.7	4.1	3.9	97.94	-2.06*	8.94	6.23	9.17†
0.199	0.21	Observed	79.2	2.9	2.7	81.79	-18.21	5.25	3.66	18.95
		Adjusted	98.8	4.4	4	100.5	0.46*	11.52	7.04	11.53†

*Denotes the smaller bias. †Denotes the smaller RMSE.

Data Analysis

The data set of weed species was collected from organic farmland located in the northern Taiwan. The record of species by 12 transect lines with length 20m each were conducted.

Table 3.
Species richness adjustment for data set of weed species, with $T = 12$, $\hat{r} = 0.82$, and $\hat{e} = 0.14$

Method	S_{obs}	Q_1	Q_2	$\hat{S}$	Estimated s.e.
Observed	74.0	19.0	9.0	92.4	11.27
Adjusted	83.6	24.1	10.6	105.4	18.68

All the details of this work can be found in my personal website by scanning the QR Code: