Anne Chao, Nicholas J. Gotelli, T. C. Hsieh, Elizabeth L. Sander, K. H. Ma, Robert K. Colwell, and Aaron M. Ellison. 2013. Rarefaction and extrapolation with Hill numbers: a framework for sampling and estimation in species diversity studies. *Ecology Monographs*.

Appendix F: Using simulation to test the proposed analytic estimators

Abundance data

We used a simulation study to compare the performance of our analytic estimators for rarefaction and extrapolation with the theoretical formulas (see Table 1 of the main text). Only the simulation results for sample-sized-based curves are presented, as all the testing procedures and conclusions for coverage-based curves are generally parallel. We simulated data from two theoretical abundance distributions and treated a large empirical diversity survey as the complete assemblage. The reference sample size was fixed at 400 for all simulations.

Data Set 1: We simulated the Zipf-Mandelbrot model in which the relative abundance takes the general form $p_i = c / (a + i)^b$, $i = 1, 2, \ldots, S$, where c is a normalized constant such that the sum of the relative abundances is 1 (Magurran 2004). Here we report the results for the special case that $S = 200$, $p_i = c / (10 + i)$, $i = 1, 2, \ldots, S$. The true (asymptotic) Hill numbers for this data set are ${}^0D = 200$, ${}^1D = 140.0$ and ${}^2D = 99.4$.

Data Set 2: We simulated the broken-stick model, with $S = 200$, $p_i = ca_i$, where c is a normalized constant and $(a_1, a_2, \ldots, a_S)$ are random variables from an exponential distribution. The true (asymptotic) Hill numbers for this data set are ${}^0D = 200$, ${}^1D = 123.2$ and ${}^2D = 91.4$.

Data Set 3: We treated the sample data of Miller and Wiegert (1989) for endangered and rare vascular plant species in the central portion of the southern Appalachian region as the true assemblage. The species-abundance distribution for this survey is given in Table F1; a total of 188 species were represented by 1008 individuals. The abundance frequency counts are: $f_1 = 61$, $f_2 = 35$, $f_3 = 18, \ldots, f_{67} = 1$. These frequencies show a relatively high degree of heterogeneity among species abundances. The true (asymptotic) Hill numbers for this data set are ${}^0D = 188$, ${}^1D = 96.9$ and ${}^2D = 54.8$.

Table F1. Abundance frequency counts of the extant rare vascular plant species (188 species, 1008 individuals) in the southern Appalachians (Miller and Wiegert 1989).

i	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
f_i	61	35	18	12	15	4	8	4	5	5	1	2	1	2	3

i	16	19	20	22	29	32	40	43	48	67
f_i	2	1	2	1	1	1	1	1	1	1

Figure F1 compares the theoretical formula and our proposed analytic estimators for rarefied and extrapolated samples up to double the reference sample size. For all three data sets, the proposed analytic rarefaction and extrapolated estimators match perfectly with the corresponding theoretical values (solid lines) for all orders in the restricted extrapolated range. However, when the extrapolated sample size is more than double the reference sample size, the performance (not shown in Fig. F1) of our predictors depends on extrapolated range and the order q. See Discussion of the main text.

Figure F1 also illustrates, for each data set, the true asymptotic value of Hill numbers. Note that the estimation target for our extrapolated estimator is the diversity for an augmented sample of finite size (the solid lines in Fig. F1), not the asymptotic value (the horizontal dashed lines). Nevertheless, in all 3 data sets, for q = 2, the extrapolated curve closely approaches the true asymptote. For q = 1, there are still some discrepancies, and for q = 0, the extrapolated species richness is substantially less than the estimated asymptote.

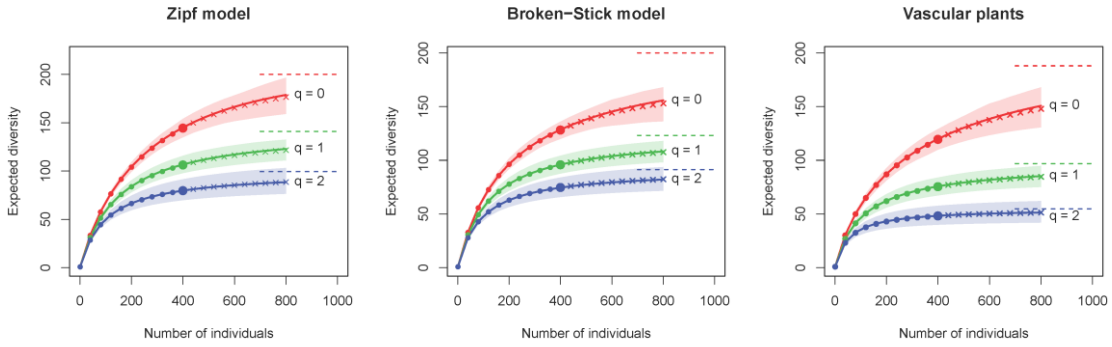

Fig. F1. Comparison of the theoretical values and analytic estimates for simulated data based on the Zipf and broken stick models and on a large empirical distribution of vascular plant abundances (in Table F1). The reference sample (larger solid dot in each curve) is set at sample size n = 400. The solid line in each sub-figure plots the theoretical formulas for ${}^0D(m)$, ${}^1D(m)$ and ${}^2D(m)$ (see Column 1 in Table 1 of the main text for formulas). Each smaller solid dot represents the average of ${}^q\hat{D}(m)$ over 200 simulated data sets for $m < 400$ (see Column 2 in Table 1 for formulas). Each "x" symbol represents the average of ${}^q\hat{D}(n+m^*)$ over 200 simulated data sets for $400 < n+m^* < 800$ (see Column 3 in Table 1 for formulas). The shaded area for each

curve represents the average 95% point-wise, unconditional confidence bands are based on 200 bootstrap replications. The horizontal dashed lines represent the true (asymptotic) Hill numbers.

Incidence data

As with the abundance data, we used a simulation study to examine the performance of our analytic estimators for rarefaction and extrapolation of incidence data (see Table 2 of the main text). We simulated data from two theoretical incidence distributions and treated a large empirical diversity survey as the complete assemblage. For all simulations, the reference sample includes data for 20 sampling units.

Data Set 1: In each sampling unit, we simulated incidence from a Zipf-Mandelbrot model with $S = 200$. The incidence probability for the ith species in any sampling unit is $\pi_i = 1 / (10 + i)$, $i = 1, 2, \ldots, S$. The true (asymptotic) Hill numbers for this data set are ${}^0\Delta = 200$, ${}^1\Delta = 141.0$ and ${}^2\Delta = 99.4$.

Data Set 2: In each sampling unit, we simulated incidence from a Broken-stick model with $S = 200$. The incidence probability for the ith species in any sampling unit is $\pi_i = 0.8\ a_i / \max(a_i)$, where $(a_1, a_2, \ldots, a_S)$ are random variables from an exponential distribution. The true (asymptotic) Hill numbers for this model are ${}^0\Delta = 200$, ${}^1\Delta = 123.2$ and ${}^2\Delta = 91.4$.

Data Set 3: In each sampling unit, the incidence probability distribution of 299 species was obtained from the incidence of 1985 tree species censused in 200 50 $\times$ 50 m quadrats in the 50 ha Barro Colorado Island (BCI) plot, Panama (Hubbell et al. 2005). The species incidence frequency counts Q_k, $k = 1, 2, \ldots, 200$ are shown in Table F2. For this data set, $Q_1 = 19$, $Q_2 = 16, \ldots, Q_{200} = 9$. The true (asymptotic) Hill numbers for this census are ${}^0\Delta = 299$, ${}^1\Delta = 189.2$ and ${}^2\Delta = 163.7$.

For all three data sets, the proposed analytic estimators for incidence data work quite well when compared with their corresponding theoretical value for rarefaction and for extrapolation up to double reference sample size (Fig. F2). The conclusions are similar to those for Fig. F1 for abundance data, and we have obtained comparable and consistent results for other species distribution models and empirical data sets.

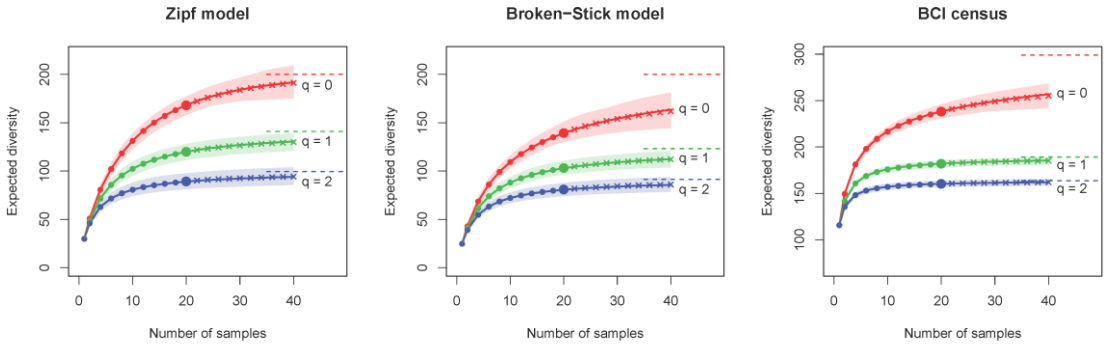

Fig. F2. Comparison of the theoretical values and analytic estimates for simulated data based on two mathematical models of species incidence and on an empirical species incidence distribution of tree census data from Barro Colorado Island (in Table F2). The reference sample (larger solid dot in each curve) represents data from $T = 20$ sampling units. The solid line in each panel plots the theoretical formulas (Column 1 in Table 2 of the main text) for ${}^{0}\Delta(t)$, ${}^{1}\Delta(t)$ and ${}^{2}\Delta(t)$. Each smaller solid dot represents the average of ${}^{q}\hat{\Delta}(t)$ over 200 simulated data sets for $m < 20$ (see Column 2 in Table 2 for formulas). Each "x" symbol represents the average of ${}^{q}\hat{\Delta}(T+t^{*})$ over 200 simulated data sets for $20 < T+t^{*} < 40$ (see Column 3 in Table 2 of the main text for formulas). The shaded area for each curve represents the average 95% point-wise confidence band are based on 200 bootstrap replications. The horizontal dashed lines represent the true (asymptotic) Hill numbers.

Table F2: Tree species incidence frequency counts determined by 200 quadrats (each of the size 50 x 50 m) from the 50 ha (1000 x 500 m) Barro Colorado Island (BCI) plot, Panama, censused in 1985 (Hubbell et al. 2005). There are 23204 incidences representing 299 species.

i	1	2	3	4	5	6	7	8	9	11	12	13	15	16	17	18	20	22	23
Q_i	19	16	9	8	3	8	4	5	6	2	3	3	1	4	1	4	4	1	2
i	24	25	28	29	30	31	32	34	35	36	37	38	39	40	41	42	43	44	45
Q_i	1	1	3	1	1	4	1	3	2	2	3	1	1	1	1	3	3	2	2
i	47	48	49	50	51	54	55	56	58	59	60	63	64	65	66	67	71	72	74
Q_i	1	2	1	1	1	1	2	5	1	1	3	4	2	1	2	1	1	2	1

i	75	82	87	88	90	92	93	94	95	98	99	100	103	104	105	107	108	111	112
Q_i	1	2	1	2	5	2	1	1	1	1	2	1	2	1	3	2	1	2	3
i	113	114	116	117	118	124	125	127	128	131	132	134	136	143	144	145	152	154	156
Q_i	1	1	1	1	1	1	1	1	2	1	2	1	1	2	3	1	2	2	2
i	158	159	161	162	165	166	167	168	170	171	172	173	177	179	181	182	184	187	188
Q_i	1	1	3	2	1	2	1	2	1	1	1	1	4	1	1	2	1	1	1
i	189	190	191	192	193	194	195	196	197	198	199	200							
Q_i	2	2	1	2	2	5	2	2	3	1	7	9							

LITERATURE CITED

Hubbell, S. P., R. Condit, and R. B. Foster. 2005. Barro Colorado forest census plot data. (http://ctfs.si.edu/datasets/bci).

Magurran, A. E. 2004. Measuring biological diversity. Blackwell, Oxford, UK.

Miller, R. I. and R. G. Wiegert. 1989. Documenting completeness, species-area relations, and the species-abundance distribution of a regional flora. Ecology 70:16-22.