Extrapolating abundance curves has no predictive power for estimating microbial biodiversity

Amy Willis, Department of Statistical Science, Cornell University

May 24, 2016

Letter to the Editor in response to Locey and Lennon (2016)

Locey and Lennon [1] recently conducted an analysis of microbial and macrobial communities to investigate the effect of sample size (N, number of individuals or reads observed) on community species richness (S), species evenness (Simpson), frequency distribution skew, and frequency count of most abundant taxa. They argue that log-log linear models fit these relationships, specifically claiming that the index of the power law between sample size and species richness is consistent across macro- and microorganisms. Furthermore, they use the "lognormal model of biodiversity" to estimate global microbial biodiversity around $10^{11}-10^{12}$ taxa. While their claims are appealing and elegant, we argue (from a statistical perspective) that their methods were inappropriate for the desired investigation.

The "lognormal model of biodiversity" (see [2] and references therein) posits that by modeling and extrapolating a species abundance curve it is possible to estimate the total number of species in the community. Unfortunately, this is untrue. Extrapolating abundance curves, accumulation curves and rarefaction curves is unsound statistical practice.

The underpinning of the invalidity of extrapolating abundance curves relates to which relationships are correlative but not predictive, versus which are correlative and predictive. Generally statisticians use independent quantities associated with the environmental system (*e.g.* pH, temperature, etc.) to predict dependent quantities (*e.g.* species richness in a lake). However, observed species richness is not a quantity associated with the environmental system: it is a result of the sampling procedure. To illustrate, consider an ecosystem comprised of bamboo, pandas, flies, and fish (true $S = 4$), and suppose our estimator of total diversity ($\hat{S}$) is sample diversity (as in [1]). We begin by only sampling $N = 20$ individuals and only observe bamboo and flies ($\hat{S} = 2$). If we continue sampling up to $N = 100$ individuals we may also find a fish ($\hat{S} = 3$), and if we continue we may eventually find a panda. However, the true number of distinct individuals in the ecosystem is unchanged for all choices of N: only $\hat{S}$ changes. In this way, while there is a correlation between N and $\hat{S}$, there is no correlation between N and S, because true biodiversity (richness) in the ecosystem exists regardless of the experiment and experimenter. In this way, the lognormal model of biodiversity has no *predictive* power for true biodiversity, only describing features of the experiment and not the universe.

The only correct (statistically admissible) way to estimate species richness is by modeling the *frequency counts*: singletons f_1, doubletons f_2, tripletons f_3, and so on. Probabilistic

models permit extrapolation from $f_1, f_2, f_3...$ to predict f_0, the number of species in the population that were not observed. The statistical literature on this problem dates to [3], with recommendations available for the best models for both macro- and microorganism richness [4, 5].

The historical popularity of extrapolating abundance curves is a poor argument for its continued use. I encourage the authors to consider the statistical perspective on this problem, and hope that improved communication between biodiversity statisticians and ecologists will advance understanding of biodiversity.

References

[1] Locey, K. J. and Lennon, J. T. (2016). Scaling laws predict global microbial diversity. *Proceedings of the National Academy of Sciences*.

[2] Curtis, T. P., Sloan, W. T., and Scannell, J. W. (2002). Estimating prokaryotic diversity and its limits. *Proceedings of the National Academy of Sciences*, **99**(16), 10494–10499.

[3] Fisher, R. A., Corbet, S., and Williams, C. B. (1943). The Relation Between the Number of Species and the Number of Individuals in a Random Sample of an Animal Population. *Journal of Animal Ecology*, **12**, 42–58.

[4] Bunge, J. and Fitzpatrick, M. (1993). Estimating the Number of Species: A Review. *Journal of the American Statistical Association*, **88**(421), 364–373.

[5] Willis, A. and Bunge, J. (2015). Estimating diversity via frequency ratios. *Biometrics*, **71**, 1042–1049.