Assessing diversity estimators and their uncertainty with checkplots

Michael Roswell, Michael Li, Rachael Winfree, Jonathan Dushoff

11/6/2020

Abstract

1. Ecologists frequently measure and compare quantities, such as Hill diversity, for which parametric uncertainty estimates are unknown or invalid. Approximate uncertainty estimates may be generated through Monte-Carlo resampling schemes such as bootstrapping. However, the bootstrapping simulation may not realize the variability in ecological metrics that occurs under random sampling of natural communities.

2. We simulated species abundance distributions with known diversity, and then sampled from them to evaluate the sampling uncertainty in Hill diversity. We then assessed whether recently proposed confidence intervals for Hill diversity estimates were valid, using two new visual diagnostic tools, the "slugplot" and the "checkplot."

3. We found that observed Hill-Simpson diversity had high sampling uncertainty even with large samples, in comparison to Hill-Shannon diversity and richness. Similarly, asymptotic Hill-Simpson diversity estimators often had higher variability than asymptotic Hill-Shannon diversity and asymptotic richness estimators, though both of these are more sensitive to rare species. The proposed confidence intervals often performed well for sample diversity, especially for more even communities. Only with very large samples could proposed confidence intervals obtain nominal coverage for asymptotic Hill-Shannon and Hill-Simpson estimators; the proposed confidence intervals performed poorly for the Chao1 richness estimator.

4. Rough estimates of relative sampling uncertainty can help guide ecologists choice of which Hill diversity to use. We found that Hill-Simpson diversity, which can be estimated with little bias, often has greater sampling uncertainty than Hill-Shannon and in some cases, richness. Slugplots and checkplots are flexible tools for assessing confidence intervals and related statistics. We showed that proposed confidence intervals may work well for sample Hill diversities, but not for asymptotic Hill diversity estimates.

Introduction

Ecologists lack precise, robust, unbiased tools to estimate the species diversity of a community based on the diversity of samples drawn from it (Haegeman et al. 2013). While ecologists know this, do they know how accurately they can estimate diversity indices? The literature conveys that richness is very hard to estimate robustly, that Simpson's index and its Hill-number version can be estimated with little bias from samples of nearly any size, and that Shannon's entropy and its Hill number equivalent fall somewhere in between (Beck and Schwanghart 2010, Chase and Knight 2013, Haegeman et al. 2013, Chao and Jost 2015). In spite of these observations, we believe that ecologists often lack intuition for how accurate and precise commonly used diversity estimates are. We suspect that ecologists tend to overestimate their ability to precisely estimate true species richness and true Hill-Simpson diversity, but underestimate their ability to describe the statistical uncertainty in the expected richness of a finite sample. As the field coalesces around the consensus that Hill diversities (eq. 1) are the best family of metrics for measuring and comparing species

diversity, identifying methods to quantify the uncertainty in Hill diversity estimates becomes more important (Willis 2019).

We define Hill diversity D as the mean species rarity in the community

$$D = \left(\sum_{i=1}^{S} p_i (r_i)^{\ell} \right)^{1/\ell}$$

where D is diversity or mean rarity, p_i is the relative abundance of species i, r_i is the rarity of species i (defined as the reciprocal of p_i), S is the total species richness, and ℓ is the scaling exponent that determines the type of mean computed (Roswell et al. ND)

Quantitative measures of the statistical uncertainty in estimates of Hill diversity (D) are technically difficult to produce, and furthermore, the existing tools rely on a variety of statistical approaches with incompatible interpretations. Thus, while a number of recent publications provide quantitative interval estimates for community diversity (Dauby and Hardy 2012, Zhang 2012, Haegeman et al. 2013, Chao and Jost 2015, Willis and Bunge 2015, Zhang and Grabchak 2016, Mao et al. 2017), the intervals they provide are not all equivalent. To the best of our knowledge, only one of these approaches is widely used in practice, namely variations of the "confidence intervals" described by Chao and Jost 2015 (Chao and Shen 2003, Colwell et al. 2012, Chao et al. 2014, 2019, Chao and Jost 2015), and included in the popular R packages SpadeR (Chao et al. 2016) and iNEXT (Hsieh et al. 2016).

The Chao and Jost 2015 confidence intervals were designed to indicate the statistical uncertainty for both sample Hill diversity and also for asymptotic Hill diversity estimates. The sampling distributions of Hill diversity estimates are poorly known, which makes defining confidence intervals for these quantities very difficult.

In this study, we describe a novel approach to assessing confidence intervals, and test the confidence intervals proposed by Chao and Jost (2015) using simulated species abundance distributions. First, we sample from simulated species abundance distributions to empirically describe the sampling distributions of Hill diversity estimates. Next, we ask whether the proposed intervals for both sample Hill diversities and for asymptotic Hill diversity estimates for the three most commonly used values for the scaling exponent ℓ (1, 0 and -1, corresponding to richness, Hill-Shannon diversity, and Hill-Simpson diversity) should be considered "confidence intervals," (Casella and Berger) using tools we introduce as "slugplots" and "checkplots."

Specifically, we ask

1. How much sampling variability is there in sample Hill diversities and in Chao and Jost's (2015) asymptotic Hill diversity estimators?
2. Does the method proposed by Chao and Jost (2015) generate valid confidence intervals that reflect this variability? If not, are these intervals consistently conservative or overconfident?
3. Are the proposed intervals from this method biased high or low?
4. Does confidence interval performance depend on the species abundance distribution or choice of Hill diversity?

STATISTICAL BACKGROUND

p-values

We define a p-value to be the probability of a particular observation (or one that is more extreme), given that a particular statistical hypothesis is true.

Formally, we define the one-tailed p-value for an observation x and the statistical hypothesis Θ as $p_\Theta(x|\Theta)$; the probability of observing a parameter as or more extreme than x, and define 1-tailed p-values as

$$p_\Theta^- = P(X \leq x|\Theta)$$

and

$$p^{+}_{\Theta} = P(X \geq x|\Theta)$$

where X is a random variable, and Θ describes the statistical hypothesis.

p-values describe an observation x in terms of the quantiles of a random variable. If Θ is true, and X is a continuous random variable, then the probability density function of $p_{\Theta}(X)$ is uniform on $[0, 1]$, by the probability integral transform (Casella and Berger 2002).

If X is a discrete random variable, however, p_{Θ^-} is not always equal to $1 - p_{\Theta^+}$. When for $x \in X$, $p^{+}_{\Theta}(x) \neq 1 - p^{-}_{\Theta}(x)$, $p_{\Theta}(x)$ is associated, in theory, with a range of "platonic" p-values between p^{+}_{Θ} and p^{-}_{Θ} [1]. When using p-values to construct confidence intervals or evaluate statistical hypotheses, it is prudent to be conservative, i.e. select only p^{+}_{Θ} or p^{-}_{Θ}, in accordance with the tail appropriate for a given hypothesis. When evaluating a p-value, it can be informative break the tie not by selecting the most conservative p-value associated with x, but instead by randomly sampling the range of p-values consistent with x. p-values estimated this way, even for discrete random variables, always have a discrete uniform distribution on $[0, 1]$.

What are confidence intervals for?

A frequentist confidence interval (CI) with confidence level L is an interval estimate obtained by a method that is expected to result in the interval containing the true value a proportion L of the time no matter what the true value is (Cox and Hinkley 1974, Casella and Berger 2002). CIs are commonly constructed as a collection of values that would not be rejected (at an aggregate p-value of $\alpha = 1 - L$) if they were treated as null hypotheses (Neyman 1937). In other words, We construct CIs using a counterfactual null hypothesis: a value falls within the CI if, imagining that that value were the true parameter value under the null hypothesis, the null hypothesis would not be rejected for the observed statistic.

Frequentist confidence intervals are commonly used to express the range of population parameter values that are consistent with the data. Thus, when valid confidence intervals can be found, they provide practical guidance about uncertainty.

Defining confidence intervals

A confidence interval denotes two limits (though one may be infinite), l and u, such that there is at least a $(1 - \alpha) * 100\%$ chance of randomly drawing a sample where $[l, u]$, computed for that sample, contains the true parameter value Θ.

$\Theta \in [l, u]$ if the statistic we observe based on the sample, $\hat{X}$, wouldn't be surprising, given that Θ is the true parameter value.

l is the lower bound when it is the minimum value satisfying $P(X \geq \hat{X}|\Theta = l) \geq a$, and u is the upper bound when it is the maximum value satisfying $P(X \leq \hat{X}|\Theta = u) \geq b$; $a + b = \alpha$

In this manuscript, we only explore equal-tailed confidence intervals, or those that have equal chances of being too high and too low. Thus, we set $a = b = \alpha/2$. Furthermore, 2-sided $(1 - \alpha) * 100\%$ confidence intervals are bounded by l and u, where l and u are given by the 1-sided $(1 - \alpha/2) * 100\%$ confidence intervals $[l, \inf)$ and $(-\inf, u]$. However, the process of imagining tests where every possible value of a parameter is the true value under the null, and finding the parameter value (l or u) that, if it were the true parameter value under the null, would result in $p_{\Theta}(x) = \alpha/2$ may be computationally or conceptually infeasible. A simpler approach is to assume a kind of symmetry such that $p_{\Theta_y}(x) = p_{\Theta_x}(y)$, where Θ_x indicates that x is the true value of the parameter of interest under the statistical hypothesis. [2]

[1] How can we cite and/or better explain this assertion?

[2] there's probably more to say about this. e.g. This assumption is implicitly true in certain problems in classical statistics (e.g. the location of a mean in a normal distribution with known variance). It seems generally like a bad assumption to me, so maybe we shoudl justify it or explain it in greater depth

Because a valid confidence interval depends on a sound method to estimate p-values (in order to conduct the counterfactual null hypothesis test), to test the validity of a confidence interval, it is sufficient to test the p-value upon which it is based. Here we introduce two diagnostic tools to assess confidence intervals and the p-values upon which they are based, which we call "slugplots" and "checkplots," respectively. A slugplot is an ordered plot of estimates and their confidence intervals, computed from Monte-Carlo random samples of $X|\Theta$. A checkplot is a histogram of estimated p-values for the simulated, true parameter value, but where p-values are computed based on a null hypothesis parameterized by a Monte-Carlo random sample of $X|\Theta$ rather than the full knowledge of Θ itself.

Slugplots

In a slugplot, confidence intervals for 1000 random samples are ordered to ease visual comparison with the true value. A slugplot makes it easy to both verify if $\alpha/2 * 100\%$ of random samples have confidence intervals above the target value and $\alpha/2 * 100\%$ have confidence intervals below the target value, and also to examine bias when nominal coverage is not achieved. A slugplot (Fig 1) is thus more nuanced than a simple measure of statistical coverage, and is useful for examining CI performance for a given combination of parameter values and sample size. Slugplots are informative for continuous statistics but may be difficult to interpret for discrete statistics.

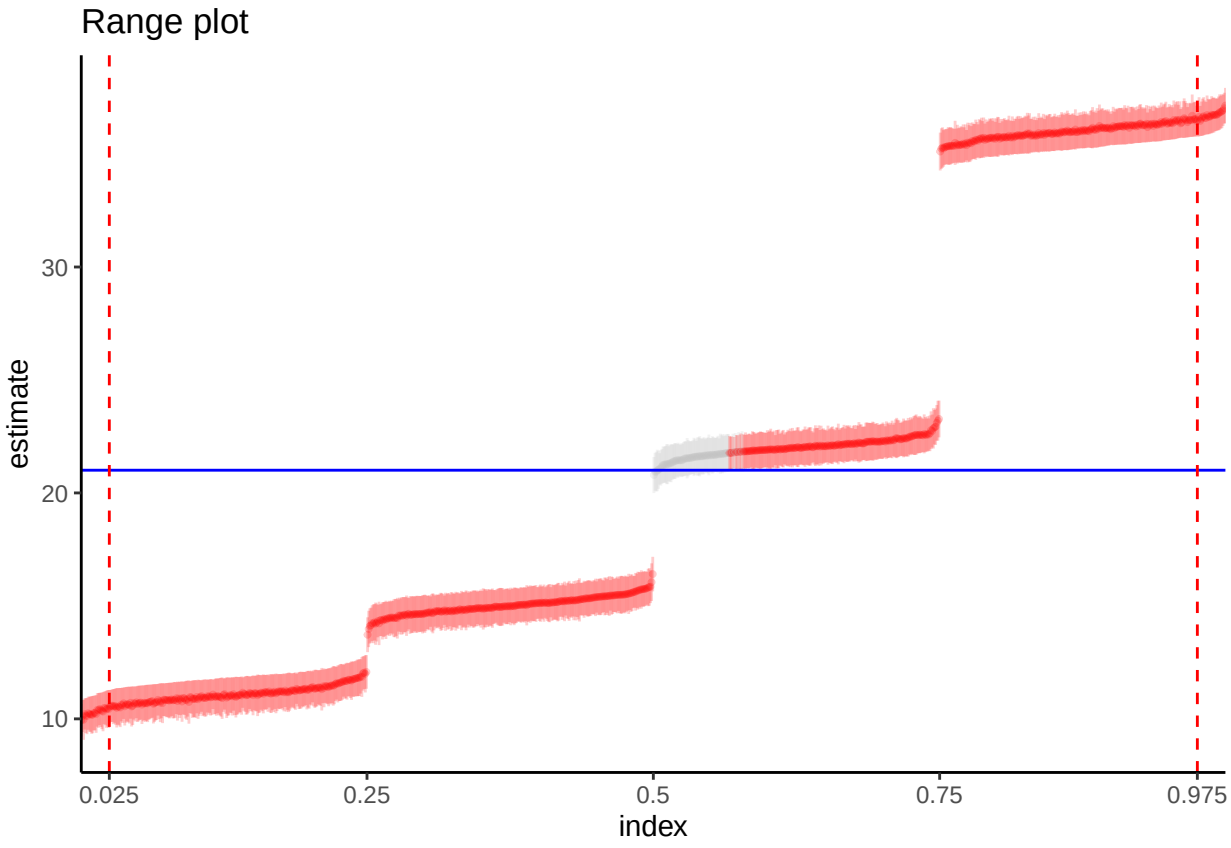

Figure 1: Figure 1. A slugplot for 95% CIs based on Student's t- type p-values, constructed for 1000 samples of 100 deviates from a normal distribution with mean= 22, standard deviation = 4. The slugplot exhibits near ideal behavior of these CIs, with 2.5% of CIs too low (red, at left) and 2.5% too high (red, at right). The point estimate for the mean for each sample is a darker shade for each CI.

Checkplots

Whereas a "slugplot" displays the empirical performance of a confidence interval (on the scale of the parameter of interest), it may be difficult to see from a given slugplot whether a confidence interval is theoretically sound. Partly, this is because it is possible to achieve nominal coverage for a particular parameter value and sample size by coincidence, despite having invalid confidence intervals. Furthermore, slugplots may be of little use for discrete statistics, for which valid confidence intervals will be conservative for most parameter values at modest sample sizes. A more general tool to assess p-values (and thus CI; a valid method to compute p-values is both necessary and sufficient to construct valid CI) is a "checkplot," a type of rank histogram.

To create a checkplot, one must first generate samples via Monte-Carlo simulation based on known parameters. From each simulated sample, one then estimates the probability that the population parameter value is less than or equal to the simulated parameter (whereas the population parameter is the simulated parameter; the probabilities are computed without considering this information). These probabilities (p-values) are then binned into a histogram (a checkplot, Fig. 2). Because we break ties for discrete statistics randomly, a valid p-value always has a flat checkplot.

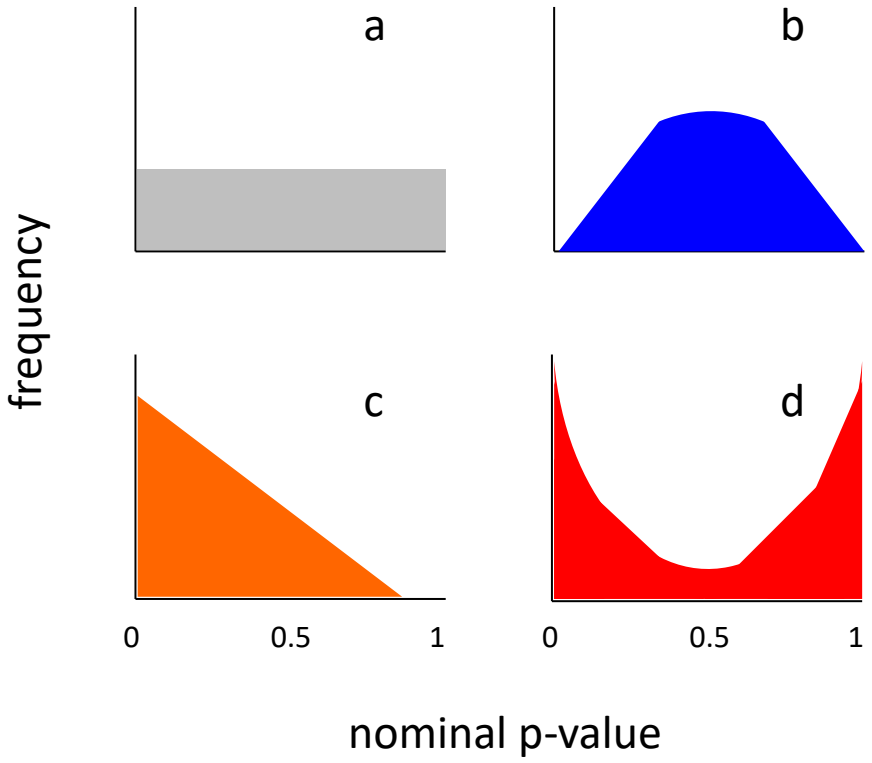

Figure 2: Fig 2. Checkplots can reveal accuracy, bias, and conservatism of a proposed p-value or confidence interval. a) an ideal checkplot has a uniform, unbroken distribution across all nominal p-values. b) checkplot with higher density around p=0.5, implying conservative confidence intervals c) Checkplot with low nominal p-values occurring more frequently than higher nominal p-values, implying biased, anti-conservative confidence intervals d) Checkplot with high density towards extreme p-values implies anti-conservative confidence intervals.

Checkplots can reveal how approximated confidence intervals or p-values may provide misleading statistical guidance ("How to interpret a p-value histogram" in press, Hamill 2001, Talts et al. 2018). Humped, centered checkplots imply conservative confidence intervals and p-values. These are obviously preferable to U-shaped checkplots, which imply anti-conservative confidence intervals and p-values. When checkplots are skewed to one side, they imply biased statistics that consistently over- or under-predict population parameters.

Because small deviations from the uniform distribution may be difficult to diagnose visually, we suggest generating up to 50,000 p-values per checkplot. Alternatively, the flatness of a checkplot might be assessed with a Chi-squared test (Wilks 2019). When developing or comparing approximated p-values, these nuanced diagnostics enable assessing CI performance beyond the simple statistical coverage probability (Fig 3).