**Cover letter**

[Greeting]

We are writing to submit [ms title] as a Letter.

The species abundance distribution (SAD) is one of the most well-known and highly-studied patterns in ecology. However, efforts to understand it in biological terms have struggled because almost all SADs – including those generated by various ecological theories - conform to the same general form, with a few common species and many rare species. One explanation for the ubiquity of this pattern is that it is a statistical artefact of the mathematical process of dividing a given number of individuals into a given number of species. However, we may still be able to detect biological influences on this pattern, if the SADs we observe in nature deviate from the shapes that would emerge statistically. Crucially, this approach hinges on there being a sufficiently well-resolved statistical baseline for us to make statistical distinctions between an observation and the most-likely shapes given the baseline – a phenomenon that may break down if the system is very small. We use combinatorics to characterize the shapes that would emerge from statistical processes for the SAD for 24,500 empirically sampled communities, and evaluate (1) whether observed SADs consistently deviate from these statistical baselines, and (2) how the specificity of the statistical baseline, and consequentially our statistical power to detect deviations between observations and the baseline, differs between very small and very large communities. We show that empirical SADs for large communities are consistently highly skewed and uneven relative to the SADs that emerge from strictly mathematical processes. However, we find that very small communities have considerably less narrowly defined statistical baselines, which greatly reduces our power to detect deviations between observed SADs and the baseline. This work highlights two novel avenues for understanding and interpreting the SAD, and ecological patterns in general. First, we are able to disentangle a biological signal from the apparent statistical constraint on the SAD, which will allow us evaluate theories in terms of how well they predict the *biological* signal in the SAD. Second, and more generally, this work directly engages with the question of to what extent we can leverage statistical baselines to interpret ecological patterns, and under what conditions this strategy breaks down.

This work goes beyond the efforts pioneered in Locey and White (2013) by leveraging the existence of the statistical constraint they document to specifically test for deviations, by explicitly exploring how the specificity of the statistical baseline, and correspondingly our ability to detect deviations, varies over large ranges in community size. We also substantially expand the range of community sizes and the number of empirical communities we analyze.

Thank you for your time and consideration.

Sincerely,