Herbivory distribution parameters in statistical models

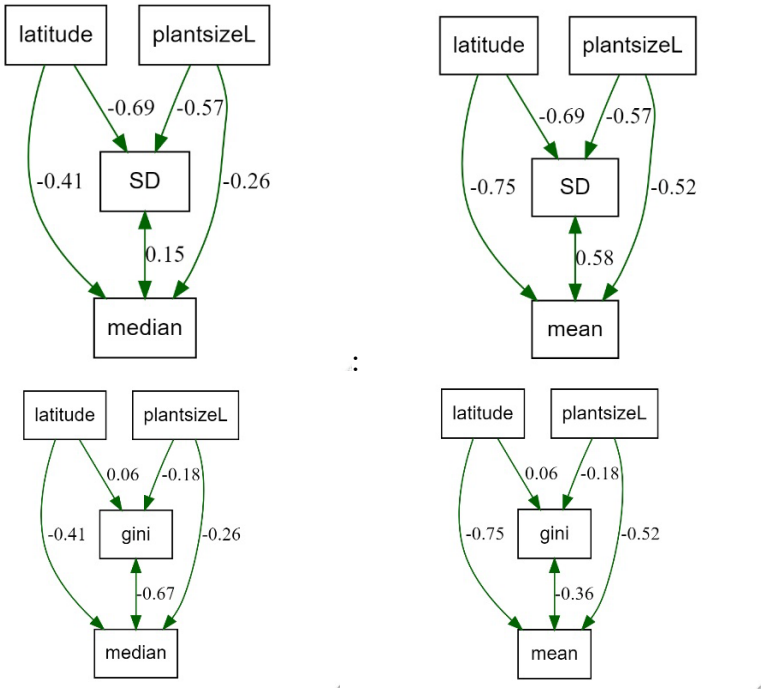
Background

Distributions are super interesting because they provide a comprehensive summary of populations or (more frequently) of samples, revealing patterns that are not immediately apparent from raw data. Parameters from distributions can be useful as both predictor and response variables in statistical models. For example, the inverse discrete Pareto distribution fits global herbivore diet breadth data, and the alpha parameter from these distributions is an excellent measure of dietary specialization for a sample. Fitting herbivory data to different distributions can be enlightening on its own, but it may also yield useful parameters for statistical models. Additionally, analyzing residuals from standard statistical models used to analyze herbivory data allows us to consider alternative inferential approaches. I plan to explore phase I herbivory data to examine fits of different distributions, considerations of the relevance of "hurdle" approaches to examining presence/absence of herbivory versus distributions of herbivory when it occurs on a leaf or plant, and causal relationships among distribution parameters, herbivore diversity, and other variables already examined (e.g., latitude).

Approach and preliminary results

We can never truly measure the exact "moments" of herbivory (or any other ecological) distributions, thus any metrics representing central tendencies, variances, skew, and kurtosis are actually latent variables from the populations or communities of interest. Approximations in sample data, such as quantiles, sums of squares, cubes, and fourth powers, can be considered "measured variables" that are influenced by these latent variables. Estimating these moments and incorporating them into statistical models is interesting to me because it helps us better understand the latent structure of the data and improve model predictions and interpretations.

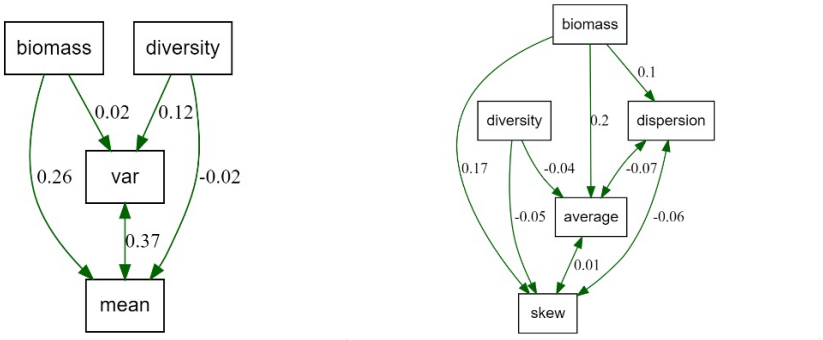
I have argued that the arithmetic mean is not the most appropriate composite variable for herbivory and the degree of positive skew should be a focal statistic to estimate. Here is an example of some SEMs comparing some Herbvar results using median and mean, sd and gini to compare the effects of latitude and plantsize on dispersion versus central tendency:



Based on this particular analysis the relationships with these measures of central tendency are as strong as the relationships with variance. But what are the best measures to use here? With *Piper* data (in a published paper) and with Phase I Herbvar data, I have estimated statistics that comprise three components of herbivory distributions: central tendency, dispersion, and skew. I've calculated quantiles (including the median) as well as the mean, geometric mean, and variance for herbivory datasets then used factor analysis of z-score transformed values of these metrics to create latent variables related to central tendency, skew, and dispersion for specialist and generalist herbivory.

I utilize a longstanding and common approach to creating latent variables: factor analysis. There are well-established justifications for such an approach, but most relevant here are: 1) dimension reduction: quantiles and other measured variables from a distribution are numerous, and factor analysis reduces the number of these variables making it easier to understand relationships between predictor and response variables; 2) reducing potential multicollinearity: factor analysis reduces multicollinearity via dimension reduction; 3) estimating latent constructs: factor analysis estimates latent constructs that may be causing the values of measured variables. This last point is important with respect to metrics that are used in statistics, such as means and standard deviations, since these do not necessarily directly measure the most relevant summaries of central tendencies or spread of the data.

Here is an example of factors for central tendency and dispersion used to examine effects of plant biomass and diversity on herbivore biomass and abundance. On the left are traditional measures (mean and variance) and on the right are three factors representing central tendency, dispersion, and skew using the same data:



The effects on dispersion and average are quite different than those on traditional measures of mean and variance. I plan to use such latent variables in Bayesian structural equation models using the most updated versions of Phase I data to test hypotheses about how herbivory distribution parameters affect and are influenced by variables such as latitude, plant size, and herbivore diversity.

Additionally, I plan to use Bayesian hierarchical models to estimate parameters of different distributions for the herbivory data, including log-normal, Weibull, and Gamma. I will test fits of these distributions at different hierarchical levels (e.g., leaf, plant, site, region) as well as examining distributions of residuals from statistical models that have been used to analyze herbivory.

Finally, I would like to explore hurdle models to compare analyses focused on presence/absence of herbivory versus levels of herbivory given that it exists.

It should all be pretty cool.