Quantifying variability in population abundances

Mario Schlemmer

Abstract

The variability in population abundances is of central concern for the quantification of evolutionary patterns and an indispensable tool for ecological analysis. A new metric proportional variability(PV), which corresponds to the expected percentage difference between abundances, is introduced here. For short time series, it is readily calculated from the differences between all combinations of abundances. A new method that simplifies the calculation for long time series is also provided. This method can also accommodate zero values if desired. The presented analysis compares PV with standard measures of population variability. The findings reveal appealing properties of this truly proportional measure and advantages over standard measures. It can provide a common ground for evaluating the variability of populations undergoing very different dynamics.

Keywords: ecological stability, extinction risk, proportional variability

Introduction

Traditionally, variability in population abundances is quantified using ad hoc measures that estimate variability from distances to the average abundance. For the coefficient of variation(CV), the mean of the original data is divided by the standard deviation. Another commonly used measure is the standard deviation of the log-transformed abundances(SDL). Recently the use of the mean abundance as the benchmark for the analysis of population variability has been questioned by researchers. Results obtained by using these traditional measures have been described as seriously flawed if the underlying population dynamics are not Gaussian, and correlations with the mean have been reported as the chief reason for pathological behavior. A clear disadvantage of these measures is that they lack a constant range, despite being relative. To overcome the perceived problem of the dependence on the average abundance and to quantify variability on a truly proportionate scale, Heith(2006) introduced a measure that quantifies variability from comparisons between all pairs of observations. It has already been applied to various contexts and, at this point, can be viewed as another standard measure of population variability.

It is given by:

$$\mathrm{H} = \frac{1}{\mathrm{C}} \sum_{comb.} \mathrm{D}(z_i, z_j), \text{ where} \tag{1}$$
$$\mathrm{D} = \begin{cases} 0 & \text{if } z_i = z_j \\ \frac{|z_i - z_j|}{\max(z_i, z_j)} & \text{if } z_i \neq z_j \end{cases}$$
$$\mathrm{C} = \frac{n(n-1)}{2}$$

Heath's population variability, referred to as H in this manuscript, is based on a total of C comparisons between abundances. It is often described as the average proportional difference between values, but it is more accurately described as the average proportionate difference relative to the higher values in a time series. In other words, if $z_i = 25$ and $z_j = 100$, the ratio of the absolute difference between values over the higher value $75/100 = 75\%$, but upon dividing the absolute difference by the lower value, this ratio $75/25 = 300\%$. Thus, adopting this new focus yields a measure that can quantify population variability as the average percentage difference relative to lower abundances

$$\mathrm{PV_S} = 1 - (1/(1 + \mathrm{MPV_S})), \text{ where} \tag{2}$$
$$\mathrm{MPV_S} = \frac{2}{\mathrm{n}^2} \sum_{comb.} \mathrm{PD_S}(z_i, z_j), \text{ where}$$
$$\mathrm{PD_S} = \begin{cases} \frac{|z_i - z_j|}{\min(z_i, z_j)} & \text{if } z_{i>0} \neq z_{j>0} \\ \mathrm{e}\frac{|z_i - z_j|}{2} & \text{if } z_i = 0 \ \& \ z_j >= 0 \end{cases}$$

Since the chronology of the data is irrelevant if this formula is used, the measures can be calculated from both ordered and unordered abundances. The unstandardized measure $\mathrm{MPV_S}$ may be defined as the expected proportional difference to higher abundances. This difference is necessarily zero for the highest value in the time series. Therefore, the sum of proportional differences is not divided through the number of pairwise comparisons C, but $\mathrm{n}^2/2$. The same logic is applicable to the variability that is proportional to higher abundances, as the difference to lower values is necessarily zero for the lowest value in a time series. The result is a differently standardized version of

H, denoted PV_I

$$PV_I = \frac{2}{n^2} \sum_{comb.} PD_I(z_i, z_j), \text{ where} \tag{3}$$

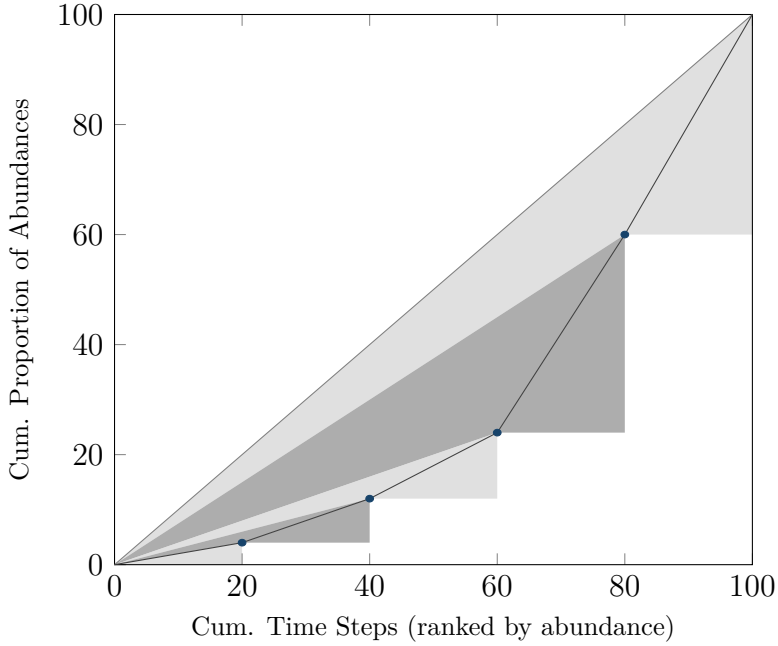
$$PD_I = \begin{cases} 0 & \text{if } z_i = z_j \\ \frac{|z_i - z_j|}{\max(z_i, z_j)} & \text{if } z_i \neq z_j \end{cases}$$

MPV_S (Eq.2) represents variability simply and clearly, but it lacks an upper bound. Following Heaths' line of thinking, a truly proportional variability measure should be standardized with a constant range. Such a metric can be obtained using the normalization operator 1-(1/(1+MPV_S). After the standardization, a score of 0.25 for PV_S indicates that the expected difference to a higher value is a third of the lower value. While the score itself still represents the difference proportional to the lower abundances, these lower abundances were previously represented by 1 but are now represented as the difference between PV_S and unity, i.e. by $1-$ PV_S. Accordingly, a score of 0.5 indicates that the expected difference between two values matches the lower value and a score of 0.90 indicates that the expected difference to a higher value is nine times larger than the lower value. If applied to the proportional difference between a single pair of abundances the result after standardization equals the value of PV_I. Although these identical scores can be viewed as expressing different perspectives, differences in the scores themselves arise only if MPV_S has been calculated from at least three abundances before standardization.

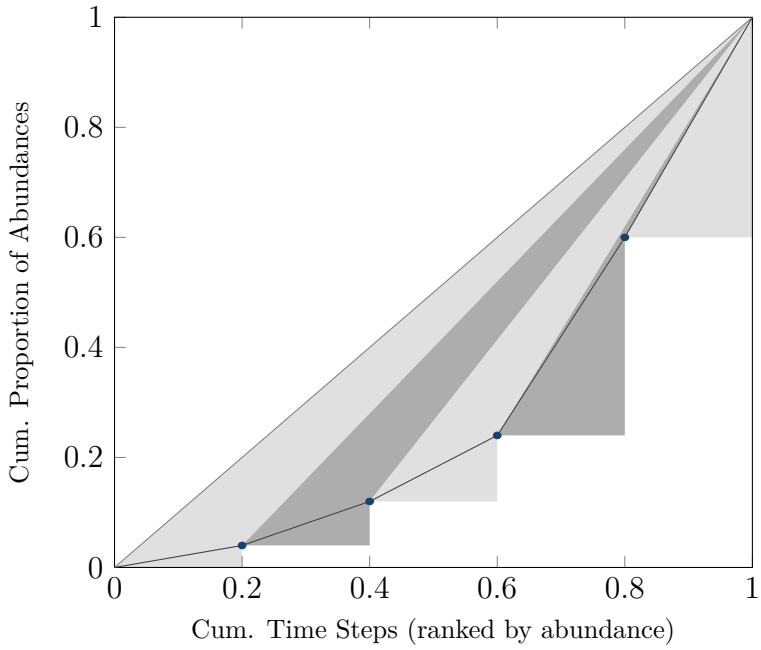
If a pairwise comparison involves a zero count, the pairwise difference PD_S is defined as the absolute difference between values multiplied with half of the base for the natural logarithm e. This avoids numerical indetermination (division by 0). An absolute difference between a zero count and positive abundance is evaluated 1.36 times higher than the same absolute difference between 1 and higher positive abundance.

Efficient Calculation

The traditional measures of variability have the clear advantage of a straightforward calculation independent of time series length. Conversely, computing proportional differences for all possible pairs of observations becomes practically impossible for longer time series without access to specialized resources.



(a) Representation of SD_I for 5 ordered abundances.



(b) Representation of SD_S for 5 ordered abundances.

Figure 1: The Lorenz curve of abundance vector (40, 80, 120, 360, 400). Each ratio between a triangle above over the corresponding triangle below a line segments equals the value of a single SD_S or SD_I in Table 1.

A new and efficient approach to calculate the proportional measures avoids comparing all pairs of abundances. Instead, each abundance is compared only once, to

obtain $\mathrm{PV_I}$ to the sum of lower values, to obtain $\mathrm{PV_S}$ to the sum of higher values

$$\begin{aligned} \mathrm{PV_I} &= \frac{2}{\mathrm{n}^2}\Big(\sum \mathrm{SD_I}\Big) \text{ , where} \\ \sum \mathrm{SD_I} &= \mathrm{C} - \Big(\sum_{i=2}^{n} \frac{\sum \min}{z_i}\Big) \end{aligned} \tag{4}$$

$$\begin{aligned} \mathrm{PV_S} &= 1 - \Big(1\Big/\Big(1 + \mathrm{MPV_S}\Big)\Big) \text{ , where} \\ \mathrm{MPV_S} &= \frac{2}{\mathrm{n}^2}\Big(\sum \mathrm{SD_S}\Big), \text{ where} \\ \sum \mathrm{SD_S} &= \Big(\sum_{i=1}^{n-1} \frac{\sum \max}{z_i}\Big) - \mathrm{C} \end{aligned} \tag{5}$$

The differentials $\mathrm{SD_I}$ and $\mathrm{SD_S}$ replace pairwise differentials $\mathrm{PD_I}$ and $\mathrm{PD_S}$, respectively, from Equation 2 and Equation 3. Their calculation is demonstrated in Table 1 using a worked example of five ordered abundances z_i. Note that for the calculation of each $\mathrm{SD_S}$ the number of pairwise comparisons c is subtracted from the ratio between the sum of higher values and value z_i, while for the calculation of each $\mathrm{SD_I}$ the ratio of value z_i over the sum of lower values is subtracted from c.

z_i	$\sum \max$	$(\sum \max)/z_i$	c	$\mathrm{SD_S}$	z_i	$\sum \min$	$z_i/\sum \min$	c	$\mathrm{SD_I}$
40	960	24	4	20	40	–	–	–	–
80	880	11	3	8	80	40	0.5	1	0.5
120	760	6.333	2	4.33	120	120	1	2	1
360	400	1.111	1	0.11	360	240	0.666	3	2.33
400	–	–	–	–	400	600	1.5	4	2.5

Table 1 Evaluation of $\mathrm{SD_S}$ and $\mathrm{SD_I}$ for 5 ordered abundances z_i.

A geometric representation of the same example is presented in Figure 1. Here abundances are represented by the areas of triangles below the line segments of the well-known Lorenz curve, while the sum of differences to lower or higher abundances for each abundance is represented by the triangles above the line segments. Therefore, each ratio between the area of a triangle above the line segment over the area of the corresponding triangle below the line segment equals the value of a single $\mathrm{SD_I}$ or $\mathrm{SD_S}$. The new technique can also be used if there are zero counts in the time series, but then it is only used for positive abundances, to which a term is added that represents the variability between positive abundances and zero counts. If $\mathrm{N_+}$ denotes the number of positive abundances in a time series and $\mathrm{N_0}$ the number of zero counts, then $\mathrm{PV_I}$ and

$\mathrm{MPV_S}$ can be calculated as:

$$\mathrm{PV_I} = \frac{2}{\mathrm{n}^2}\Big(\sum \mathrm{SD_{I_+}} + \sum \mathrm{SD_{I_0}}\Big) \text{ , where} \tag{6}$$

$$\sum \mathrm{SD_{I_0}} = \mathrm{N_0} \cdot \mathrm{N_+}$$

$$\mathrm{MPV_S} = \frac{2}{\mathrm{n}^2}\Big(\sum \mathrm{SD_{S_+}} + \sum \mathrm{SD_{S_0}}\Big) \text{ , where} \tag{7}$$

$$\sum \mathrm{SD_{S_0}} = \mathrm{N_0} \cdot \frac{\mathrm{e}}{2}\sum_{i=1}^{n} z_i$$

Dual Measurement of Proportional Variability

The indices $\mathrm{PV_I}$ and $\mathrm{PV_S}$ are both proportional measures of variability, each with a distinct focus. While the former quantifies variability proportional to higher abundances, the latter quantifies variability proportional to lower abundances. Combining the two yields an index accuratly denoted as proportional variability(PV)

$$\mathrm{PV} = 1 - \Big(1\Big/\Big(1 + \mathrm{MPV}\Big)\Big) \text{ , where} \tag{8}$$

$$\mathrm{MPV} = \frac{1}{\mathrm{n}^2}\sum_{comb.}\left(\left(\frac{|z_i - z_j|}{\max(z_i, z_j)}\right) + \left(\frac{|z_i - z_j|}{\min(z_i, z_j)}\right)\right) = \frac{\mathrm{MPV_S} + \mathrm{PV_I}}{2}$$

For the calculation of PV, the number of comparisons doubles to n(n-1), which means the sum of proportional differences has to be divided by n^2 to obtain the value of the mean proportional variability MPV, which is standardized in the usual way. PV expresses the expected average percentage difference to other abundances. This new metric accounts for both the mutually corresponding proportional differences for each pair of abundances, instead of quantifying population variability only relative to either lower or higher values. Both aspects of proportional variability have a marked influence on the value of PV for low and medium levels of variability. Nonetheless, the influence of $\mathrm{MPV_S}$ on the value of PV increases with variability, as only the value of $\mathrm{PV_I}$ is bounded by 1.

Another way to utilize the proportional differences from which $\mathrm{PV_I}$ and $\mathrm{PV_S}$ are calculated corresponds to the average difference between their pairwise differentials. If the extensive method of pairwise comparisons is used differentials can be either substracted or multiplied to calculate $\mathrm{MPV_D}$. If the efficient method outlined above is used the

smaller value of PV_I (Eq.4) is subtracted from MPV_S (Eq.5) before standardization.

$$\begin{aligned} PV_D &= 1 - \left(1 \Big/ \left(1 + MPV_D\right)\right) \text{ , where} \\ MPV_D &= \frac{2}{n^2} \sum_{comb.} \left(\left(\frac{|z_i - z_j|}{\max(z_i, z_j)}\right)\left(\frac{|z_i - z_j|}{\min(z_i, z_j)}\right)\right) \\ &= \frac{2}{n^2} \sum_{comb.} \left(\left(\frac{|z_i - z_j|}{\max(z_i, z_j)}\right) - \left(\frac{|z_i - z_j|}{\min(z_i, z_j)}\right)\right) = MPV_S - PV_I \end{aligned} \tag{9}$$

PV_D deemphasizes small differences between pairs of abundances (see Figure 2), which makes the overall score more sensitive to remarkably low values in a time series. This is the opposite of how PV_I and H evaluate differences. The convex function of PV_D for pairwise differentials intersects the function of PD_I if abundance z_j is twice as large as abundance z_i at a score of 0.5 ($PD_I = 0.5$, $PD_S = 1$), and the function of PD if abundance z_j is three times larger than abundance z_i, at a score of 1.33 ($PD_I = 0.66$, $PD_S = 2$).

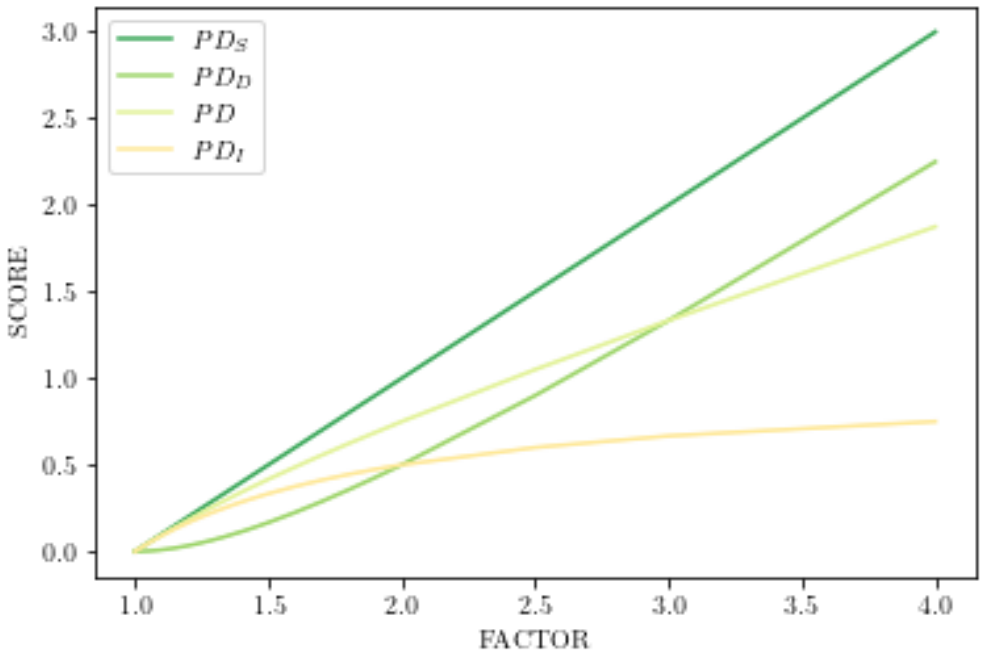


Figure 2: Comparison of pairwise difference metrics PD_S, PD_D, PD and PD_I, in terms of their evaluation of low and medium levels of variability between pairs of abundances(n=2).

Extended and Angular Proportional Variability

The proportional approach to variability measurement can be extended by combining proportional and absolute differentials between abundances.

$$\mathrm{DDI} = \frac{1}{\mathrm{n}^2\mu}\sum_{comb}\left(\frac{|z_i - z_j|}{\max(z_i, z_j)}|z_i - z_j|\right) = \frac{1}{\mathrm{n}^2\mu}\sum_{comb}\frac{|z_i - z_j|^2}{\max(z_i, z_j)} \tag{10}$$
$$\mathrm{DDI}^\delta = \frac{1}{\mathrm{n}^2\mu}\sum_{comb}\left(\frac{|z_i - z_j|}{\max(z_i, z_j)}\right)^\delta |z_i - z_j|$$

The dyadic disparity index(DDI) is obtained by multiplying the $\mathrm{PD_I}$ (Eq.3) for each pair of abundances with the absolute difference between them. It can also be generalized by potentiating each $\mathrm{PD_I}$ with parameter δ>0 before it is multiplied with the absolute difference. A DDI with a parameter of $\delta = 2$ is more sensitive to left-skewed distributions than the standard DDI, while a DDI with $\delta = 0.5$ is more sensitive to right-skewed distributions.

A measure very similar to H is denoted as angular proportional variability(APV), and can be written as:

$$\mathrm{APV} = \frac{2}{C\pi}\sum_{comb}|atan2(z_i, z_j) - atan2(z_j, z_i)| \tag{11}$$

APV is based on the difference between two angles from the positive x-axis to two coordinates of the abundance pair in the unit square. The two coordinates for each pairwise comparison are obtained by drawing each value once at the x-axis, while the other value is drawn at the y-axis. The angles are calculated using the function arctan2. Both DDI and APV lack the ease of computation of CV and H, the measures they are most similar to. Even though for APV, the accommodation of zero values is straightforward in most programming languages, the computation of H is overall more efficient if calculated with the new approach outlined above.

Comparison of Different Measures

The new proportional measures $\mathrm{PV_S}$, PV and $\mathrm{PV_D}$ share similarity with the time-honored measure SDL, particularly for the evaluation of variability for non-Gaussian population dynamics and cases involving zero counts(see Table 2). Nonetheless, to calculate SDL for time series y4 and y5, the constant 1 had to be added to all abundances(log z+1) as log(0) is undefined. It has long been known that this leads to a bias whenever there are low abundances regardless of the variance-mean relationship(Gaston & McArdle, 1994). There is no bias for the new proportional measures, but similar to SDL they indicate high variability if there is a smaller percentage of

zero counts in a time series and values for positive abundances are high, while a high percentage of zero counts is correctly identified as low variability especially if the values for the positive abundances are low. Note that CV acts more like a concentration index, and H evaluates each difference between zero counts and positive abundances with the highest possible difference, irrespective of the magnitude of the difference.

y	DDI	CV	H	PV	SDL	PV_S	PV_D
$\mathbf{y_1}$=(1, 7, 9, 9, 10, 11, 11, 12, 13, 17)	0.122	0.416	0.374	0.546	0.786	0.674	0.634
$\mathbf{y_2}$=(5, 5, 5, 6, 7, 8, 13, 16, 17, 18)	0.162	0.540	0.418	0.399	0.540	0.488	0.366
$\mathbf{y_3}$=(5, 5, 5, 5, 6, 7, 7, 8, 10, 42)	0.296	1.136	0.357	0.451	0.647	0.570	0.501
$\mathbf{y_4}$=(0, 0, 1, 2, 12, 15, 16, 16, 19, 19)	0.372	0.822	0.666	0.815	1.276	0.891	0.884
$\mathbf{y_5}$=(0, 0, 0, 0, 0, 0, 0, 0, 1, 2)	0.817	2.250	0.366	0.334	0.415	0.402	0.255

Table 2 Example abundance time series

What follows is an analysis of the new measure PV that builts on and partly replicates a previous comparison between different measures(Heath, 2006). All four measures included in the presented analysis take on similar values if abundances in a time series are normally distributed (see Figure 3).

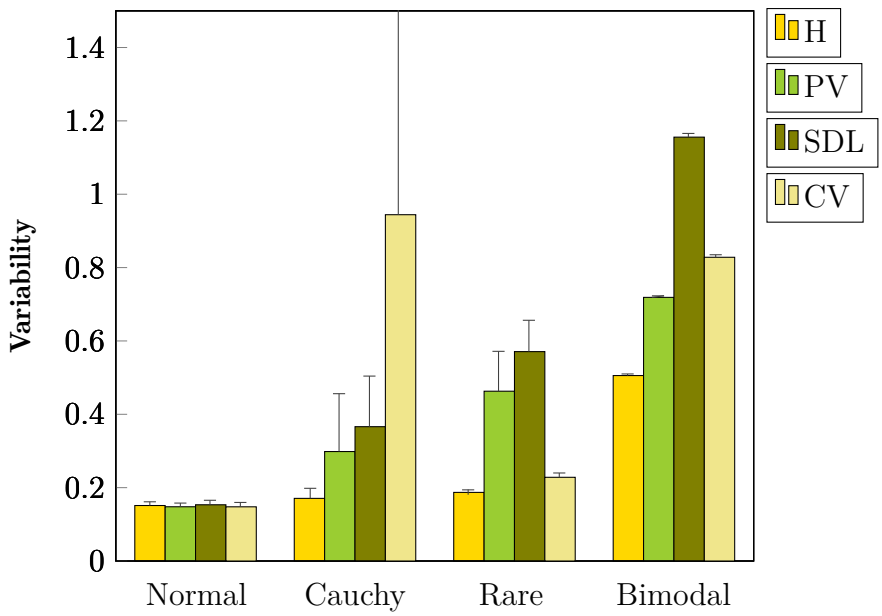


Figure 3: Comparison of H, PV, SDL, and CV across 4 simulations of population change for time series of 100 intervals. Each simulation was conducted 100.000 times. For the first simulations, populations were generated by randomly drawing abundance from a Gaussian distribution with a mean of 1000 and a standard deviation of 150. Absolute values from the heavy-tailed Cauchy distribution (location 100, scale parameter 5) were used for the comparison denoted as *Cauchy*. Rare events were simulated by selecting 95 abundances from a stable distribution (mean=100; SD=5), but letting the population crash at a frequency of 5%, by drawing values randomly from a uniform distribution ranging from 1 to 20. Values for the final comparison denoted as *Bimodal* were calculated with mean 100(SD=10) for 50 years and mean 10(SD=1) for the remaining 50 years. Error bars are standard deviations.

The heavy-tailed Cauchy distribution was used to illustrate how the measures react to one or several population crashes and bonanzas in a time series. Bonanzas are rare population booms.For individual time series in the *Cauchy* condition PV, SDL and CV are more responsive if there are abundances that are orders-of-magnitude larger or smaller than most other abundances. The high values of CV are due to bonanzas and appear excessive, PV and SDL are more sensitive if populations crash. The value of H does not exceed 0.3 for any time series in the *Cauchy* condition. A third condition illustrates a population crash at a frequency of 5%, here CV and H have been described as being more robust. Viewed differently they lack sensitivity to population crashes, a property that renders them unsuitable for crucial areas of ecological analysis, in particular, the evaluation of extinction risk and the identification of demographically and genetically important populations.

All four measures indicate relatively high variability for the *Bimodal* condition that illustrates the effects of sudden shifts in carrying capacity. This can illustrate either a scenario with longer periods of stability between population shifts following changes in the environment or more frequent shifts corresponding to cyclical dynamics(Heath, 2006). In either scenario, the increased values of variability seem appropriate, as half of the pairwise comparisons are between abundances that are, on average, an order of magnitude apart. Note that at this level of variability H has already reached its upper bound of about 0.5 for non-exponential population dynamics (Heith & Borowski, 2013). Even if the difference between the two carrying capacities would increase by additional orders of magnitude, the value of H would hardly increase any further. This upper bound can be higher than 0.5 for shorter time series as H is not replication-invariant. If a given time series is cloned such that the new times series is twice as long as the original, H indicates lower variability. Not only does this inconsistency make population variability dependent on the length of the time series, but also the frequency of measurement. Irrespective of the length of the time series, in many cases, H has no potential to increase any further if non-exponential variability increases, although the value of H is only at a level of about 0.5. In contrast, PV increases from 0.72 to 0.96 if the difference between the carrying capacities increases by another magnitude. Although biological systems exhibit large variations in behavior, owing to environmental constraints such as limited resources and predation, the theoretical probability of observing exponential population dynamics in a time series is low. This suggests that PV makes more effective use of its range from 0 to 1, resulting in a more intuitive score and more prominent differences. PV also satisfies the mathematical property of replication invariance.

The correlation between the measures and the average deviation from the mean(AMD) were also analyzed. Correlations with ADM have been previously reported for CV and SDL but not H if abundances are randomly drawn from a heavy-tailed Cauchy distri-

bution(Heath, 2006). Based on this apparent robustness to non-Gaussian behavior, H has been recommended as a measure that can provide a common ground for the summarization and comparison of variability. The analysis presented here also includes the correlations for the *Rare* and *Bimodal* conditions and comes to a different conclusion. As expected, all measures are correlated with ADM if values are normally distributed(see Figure 4). Although the previous analysis reported a high correlation between CV and ADM for the heavy-tailed *Cauchy* condition only a moderate correlation was found in the current analysis.

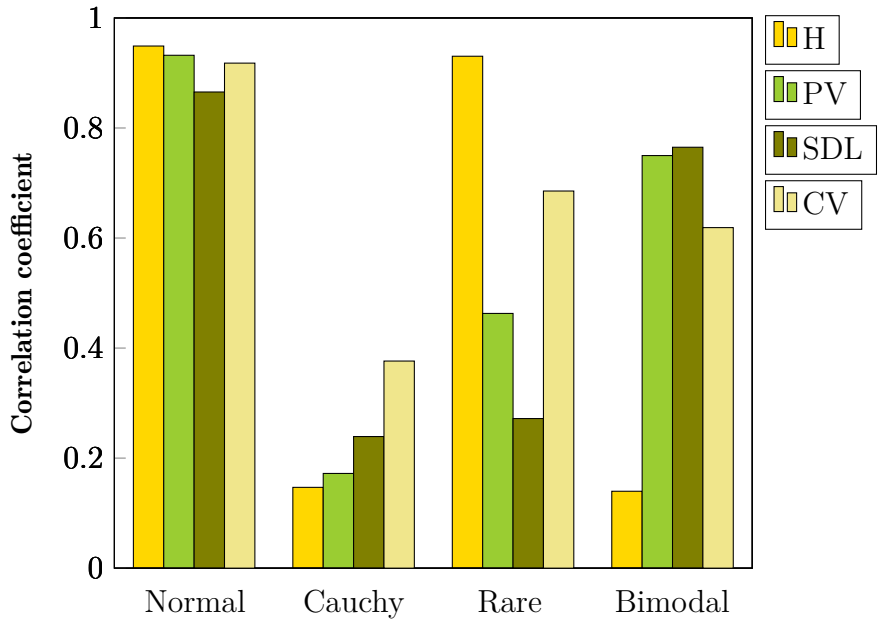


Figure 4: Correlation of H, PV, SDL and CV with average deviation from the mean (ADM), across 100.000 simulations of time series with 100 intervals. See caption of Figure 3 for details on the 4 conditions.

In the *Rare* condition, the variation in the value of H occurs because of proportionally small differences between values during the long stable periods. This leads to a high correlation between H and ADM as the larger number of proportionally small differences has a greater effect on ADM. In contrast, PV and SDL are more sensitive to the extent of the population crash, making them more independent of ADM in the *Rare* condition. In the *Bimodal* condition, H is more sensitive to the variation within the vastly different carrying capacities than to the variation between them. The result is a lower correlation between H and ADM as ADM is now associated with the difference between the two carrying capacities. The *Bimodal* condition demonstrates again that H loses sensitivity if differences between abundance pairs are already large and increase further, and it only remains sensitive to the proportionally small changes, often of less environmental concern. The findings also indicate that all four measures have a strong relationship with the mean abundance for certain types of dynamic be-

haviors. However, this should not be concluded as problematic, just because CV shows pathological behavior in the *Cauchy* condition. Instead, the focus should be on the properties of variability measures that are important for population dynamics from a conservation perspective, particularly sensitivity to population crashes and long-term population decline. For this primary objective, only SDL and PV are appropriate, but between SDL and PV it is the novel measure that has the advantages of a constant range, a more intuitive interpretation, and unbiasedness when there are zero counts in the time series. On this basis, PV is likely the best option to provide a common ground for the quantification of variability in a wide range of ecological contexts.

Concluding Remarks

Population variability is an important concept in biological theory and practice. In this manuscript, a new measure, *proportional variability* (PV) is introduced that is calculated from the proportional differences between all possible pairs of abundances in a time series. It is closely related to a previously described measure that quantifies variability relative to the higher values(Heith, 2006). It also accounts for proportional differences relative to lower values, which gives it the intuitive interpretation as the expected percentage difference to other abundances. Other measures are also introduced here that can be calculated using this proportional approach, and a new method is described, which avoids the comparison of all abundance pairs. PV is simple to use even for long time series and possesses desirable properties, it can provide a common ground for evaluating the variability for a remarkable range of dynamical patterns of abundance. In contrast, CV and H are inadequate as measures of population variability from the perspective of conservation ecology considering global environmental trends, as illustrated by their marked insensitivity to population crashes. They may still have a place in other contexts. SDL has also some shortcomings. From the other novel measures introduced here, some are similar to PV in many aspects, but only PV is suggested to be further investigated as a measure of variability in the field of ecology, to promote comparison. The two other new proportional measures PV_S and PV_D can be useful in other disciplines to study the relationship between scalar quantities. Both these measures satisfy the transfer principle, which makes them appropriate for the study of income and wealth distributions. If a progressive transfer occurs between observations they indicate lower inequality. PV_D goes one step further and is also transfer sensitive (Kolm, 1976). Given two pairs of observations that have the same absolute difference between them in an ordered distribution, then the value of PV_D decreases more if the progressive transfer occurs between the two lower-ranked observations. These measures can also be adapted to a wide range of disciplines throughout the sciences.

References

Gaston, K. J., & McArdle, B. H. (1994). The temporal variability of animal abundances: measures, methods and patterns. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences, 345*(1314), 335-358.

Heath, J.P. (2006). Quantifying temporal variability in population abundances. *Oikos, 115*(3), 573-581.

Heath, J.P., & Borowski, P. (2013). Quantifying proportional variability.*PLoS One, 8*(12), e84074.

Kolm, S. C. (1976). Unequal inequalities. II. *Journal of Economic Theory, 13*(1), 82-111.