Quantifying variability in population abundances

Mario Schlemmer

The variability in population abundances is of central concern for the quantification of evolutionary patterns and an indispensable tool for ecological analysis. Standard measures of population variability are often biased, insensitive to population crashes, and exhibit other pathological behavior. Here I introduce new measures based on the proportional difference between all pairs of abundances and describe a computationally efficient method for their calculation. The new measures are compared to standard measures of population variability. The findings reveal that they have appealing properties and advantages over standard measures. They can provide a common ground for evaluating the variability of populations undergoing very different dynamics.

Keywords: ecological stability, extinction risk, proportional variability, dispersion

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 popular 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 inaccurate if the underlying population dynamics are not Gaussian, and correlations with the mean have been reported as the chief reason for pathological behavior. Nonetheless, the focus of the two traditional measures is very different, particularly for distributions that are not Gaussian. A clear disadvantage of these measures is that they lack a constant range, despite being relative. To address 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 can quantify variability from comparisons between all pairs of abundances. This measure has already been applied to various contexts. It is given

by

$$\mathrm{PV}^- = \frac{1}{\mathrm{C}} \sum_{comb.} \mathrm{D}^-(z_i, z_j), \text{ where}$$

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

$$\mathrm{C} = \frac{n(n-1)}{2}$$

The measure is based on a total of C comparisons between abundances and is often described as the average proportional difference, but it only measures the average difference relative to the higher values in each pair of abundances. If $z_i = 25$ and $z_j = 100$, the ratio of the absolute difference between values over the higher value $75/100 = 75\%$. But if the same absolute difference is divided through the lower value, this ratio $75/25 = 300\%$. Calculated for all pairs of abundances yields a metric that quantifies proportional variability relative to lower values. Furthermore, both ratios can be combined resulting in a single measure of proportional variability.

Proportional Variability

The average proportional difference relative to lower abundances is defined as:

$$\mathrm{PV}^+ = \frac{1}{\mathrm{C}} \sum_{comb.} \mathrm{D}^+(z_i, z_j), \text{ where}$$

$$\mathrm{D}^+ = \begin{cases} \frac{|z_i - z_j|}{\min(z_i, z_j)} & \text{if } z_{i>0} \neq z_{j>0} \\ \mathrm{e}\frac{|z_i - z_j|}{2} & \text{if } z_i = 0 \ \& \ z_j >= 0 \end{cases}$$

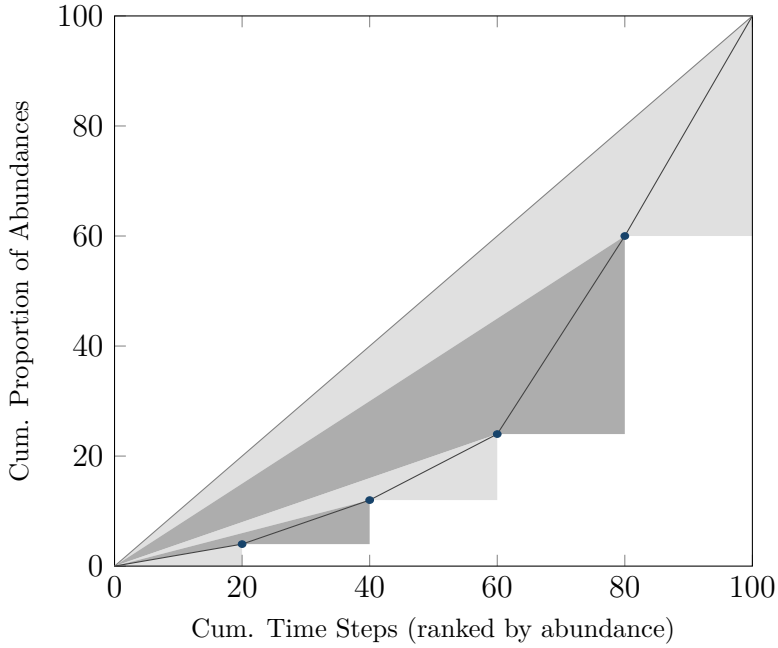
Based on all pairwise comparisons of abundances in a time series, PV^+ expresses the average proportion of the absolute differences relative to the lower abundances. The lower bound of the metric is zero and unlike PV^- it has no upper bound. A standardized measure is given by:

$$\mathrm{PV}' = 1 - (1/(1 + \mathrm{PV}^+))$$

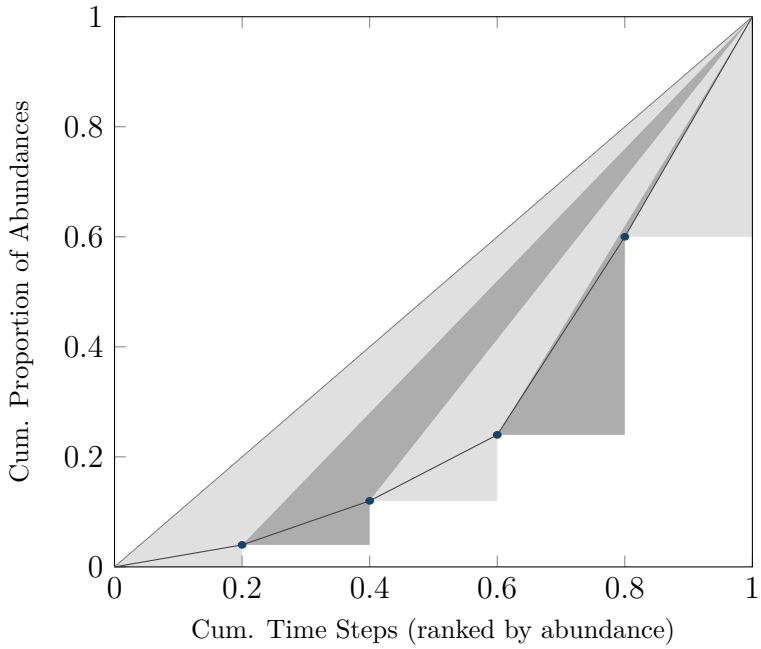
$$\mathrm{PV}'' = 1 - (1/(1 + \mathrm{PV}^-))$$

Bound between 0 and 1, the metric PV^+ still represents variability on a proportional scale. A score of 0.25 indicates that the average 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' and unity, i.e. by $1-\ \mathrm{PV}'$. 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. The same standardization applied to PV^- yields PV''.



(a) Representation of differentials PD^- for 5 ordered abundances.



(b) Representation of differentials PD^+ for 5 ordered abundances.

Figure 1: The Lorenz curve of abundance vector (40, 80, 120, 360, 400). The ratios between the triangles above over corresponding triangles below the line segments equal the values of differentials PD^+ and PD^- in Table 1.

If a pairwise comparison involves a zero count then the difference D^+ is defined as the absolute difference between values multiplied with half of the base for the natural logarithm($e = 2.7182...$). 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 abundances.

The standard ad hoc measures of variability CV and SDL have the clear advantage that their calculation is computationally efficient, even for long time series. The standard approach under which truly proportional metrics have been developed is not nearly as straightforward. As every single pair of abundances is compared the computation can be a very tedious task and for longer time series it is practically impossible without the proper resources. A new and efficient approach to calculating the proportional measures avoids comparing all pairs of abundances. It is based on calculating only one proportional difference to higher(or lower) values for each abundance. If abundances are sorted from lowest to highest, PV^+ is then given by:

$$PV^+ = \frac{1}{C}\left[\left(\sum_{i=1}^{n-1} \frac{\sum \max}{z_i}\right) - C\right]$$

and PV^- by:

$$PV^- = \frac{1}{C}\left[C - \left(\sum_{i=2}^{n} \frac{\sum \min}{z_i}\right)\right]$$

The calculation of proportional differences towards all higher or lower values(denoted PD^+ and PD^-) for each abundance is demonstrated in table 1 using a worked example.

Z	$\sum \max$	$(\sum \max)/z_i$	c	PD^+	Z	$\sum \min$	$(\sum \min)/z_i$	c	PD^-
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 PD^+ and PD^- for time-series Z.

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 Lorenz curve, while the sum of differences are represented by the triangles above these 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 PD^+ or PD^-.

This efficient method to calculate proportional variability can also be used if there

are zero counts in the time series. In that case the above formula is used to calculate the proportional differences between positive abundances, then a term is added that represents the variability between positive abundances and zero counts.

$$\sum \mathrm{PD}^{-}_{z=0} = \mathrm{N}_{z=0} \cdot \mathrm{N}_{z>0}$$
$$\sum \mathrm{PD}^{+}_{z=0} = \frac{\mathrm{e} \cdot \mathrm{N}_{z=0}}{2} \sum_{i=1}^{n} z_i$$

The number of zeros in the distribution either has to be multiplied with the number of positive abundances, or it is multiplied with half of the constant e and the sum of abundances. The two proportional variabilities relative to higher and lower abundances can be combined to a single index of proportional variability(PV):

$$\mathrm{PV} = \frac{\mathrm{PV}' + \mathrm{PV}''}{2}$$

PV reflects the perspectives from lower and higher abundances, taking the mean makes sure that both the average proportional differences to lower and higher values contribute equally to the variability index.

Comparison of Indices

Indices quantify variability very differently when the distribution of abundances is not Gaussian. The indices also differ in how they accommodate zero counts. Both CV and PV^{-} indicate identical variability if a time series with zero counts is scaled by a factor, while the values of SDL and PV^{+} change.

y	CV	SDL	PV^{-}	PV^{+}	PV''	PV'	PV
$\mathbf{y_1}$=(1, 7, 8, 9, 10, 11, 12, 13, 14, 15)	0.41	0.79	2.32	0.39	0.70	0.28	0.49
$\mathbf{y_2}$=(5, 5, 6, 6, 7, 7, 8, 8, 8, 40)	1.06	0.60	1.25	0.32	0.56	0.24	0.40
$\mathbf{y_3}$=(5, 5, 8, 8, 10, 10, 12, 12, 15, 15)	0.36	0.40	0.67	0.34	0.40	0.25	0.33
$\mathbf{y_4}$=(0, 7, 8, 9, 10, 11, 12, 13, 14, 16)	0.45	0.72	5.86	0.41	0.85	0.29	0.57
$\mathbf{y_5}$=(0, 0, 1, 1, 1, 1, 2, 2, 2, 2)	0.66	0.42	1.08	0.55	0.54	0.35	0.43
$\mathbf{y_6}$=(0, 0, 1, 3, 10, 12, 15, 17, 20, 22)	0.85	1.26	11.73	0.70	0.92	0.41	0.67
$\mathbf{y_7}$=(0, 0, 0, 0, 0, 10, 10, 20, 30, 30)	1.25	1.59	17.77	0.65	0.95	0.39	0.67
$\mathbf{y_8}$=(0, 0, 0, 0, 0, 1, 1, 2, 3, 3)	1.25	0.60	1.78	0.65	0.64	0.39	0.52
$\mathbf{y_9}$=(0, 0, 0, 0, 0, 0, 0, 0, 1, 99)	3.28	1.45	23.95	0.38	0.96	0.27	0.62
$\mathbf{y_{10}}$=(0, 0, 0, 0, 0, 0, 0, 0, 1, 2)	2.25	0.39	0.68	0.37	0.40	0.27	0.34

Table 2 Evaluation of variability for sorted example time series.

This sensitivity to the context in which zero counts occur can be viewed as an advantage. The probability of zero counts in a time series is higher if abundances are low. A metric should therefore indicate lower variability if zero counts are observed in a time series with low abundances, compared to a time series in which the population drops to zero from higher abundances. SDL has this property, but to calculate SDL if there are zero counts the data has to be transformed, as log(0) is undefined. The constant 1 is added to each abundance. 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). The proposed measures don't share this bias.

What follows is an analysis that compares the scores of existing measures and the proposed measure PV for different population dynamics. This partly replicates a previous comparison between measures by 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).

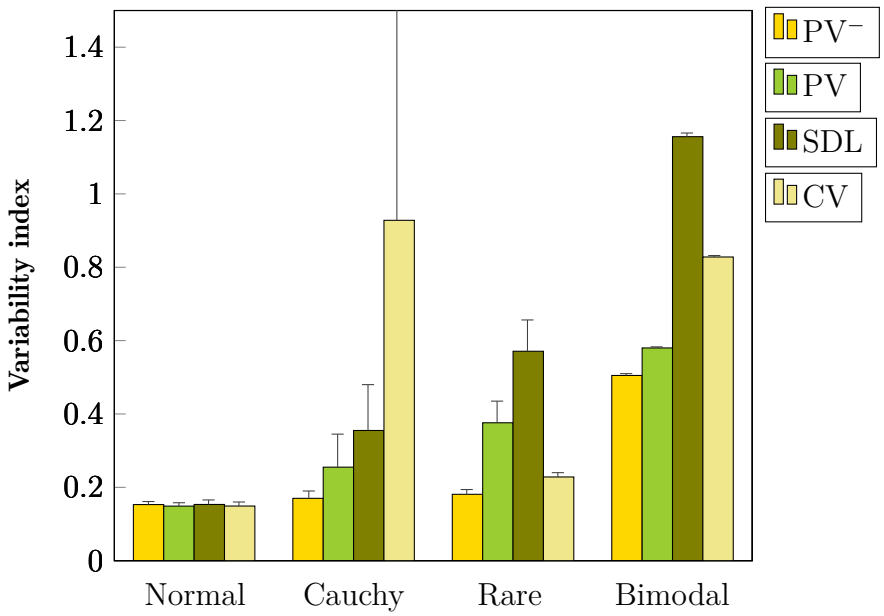


Figure 2: Comparison of PV^-, PV, SDL, and CV across 4 simulations of population change for time series of 100 intervals. Each simulation was conducted 50.000 times. For the first simulations, populations were generated by randomly drawing abundance from a Gaussian distribution with a mean of 1000 and a standard deviation of 150. The heavy-tailed Cauchy distribution (location 100, scale parameter 5, restricted to natural numbers) was used to simulate extreme outliers. 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 *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. The very high variability(mean $= 0.93, \mathrm{SD} = 1.31$) of CV in this

condition are the result of extremely high abundances at one or two time steps. CV is very sensitive to high values while PV^- is the most robust measure in this condition. A third condition illustrates a population crash at a frequency of 5%, here PV and SDL are sensitive while CV and PV^- aren't. This lack of sensitivity makes them unsuitable for crucial areas of ecological analysis, in particular, the detection of long-term population decline, the evaluation of extinction risk, and the identification of demographically and genetically important populations. For these primary objectives, SDL, PV^+, and PV are better alternatives.

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 PV^- 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 PV^- would hardly increase any further if non-exponential variability increases. For this measure to take on higher values, a substantial proportional difference between nearly all observations is necessary. 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 over many time steps is low. In contrast, PV' increases from 0.72 to 0.96 if the difference between the carrying capacities increases by another magnitude. As a consequence PV, which is based on both types of proportional variability, also makes more effective use of its range from 0 to 1.

The discussed variability indices aren't affected by the temporal order of abundances. A measure of proportional variability can also be calculated only for consecutive time steps

$$\mathrm{CD} = \frac{1}{n-1}\sum_{cons}\left(\frac{|z_i - z_j|}{2z_i} + \frac{|z_i - z_j|}{2z_j}\right) = \frac{1}{n-1}\sum_{cons}\left|\frac{z_i^2 - z_j^2}{2(z_i \cdot z_j)}\right|$$

Recently Fernández-Martínez et al.(2018) proposed a similar measure based on the natural logarithm of the ratio x_i/x_j of consecutive values. The new proportional measure CD is more sensitive to remarkably low values compared to the previouslsy described measure of temporal variability. It can also be adapted to accommodate zero counts without bias. As the differences between abundances that are not consecutive do not contribute to variability, such a measure of temporal variability would lead to very different variability scores for the two scenarios of the *Bimodal* condition described

above.

Variability measures also differ in how strongly they are affected by the average deviation from the mean. The mean is strongly influenced by outliers and a strong correlation is therefore a sign that a measure may not be appropriate, particularly for non-Gaussian population dynamics. I calculated the correlation between the measures and the average deviation from the mean. The Bravais-Pearson correlation coefficient was calculated for the *Normal*, *Rare* and *Bimodal* conditions, Spearman's rank correlation coefficient for the *Cauchy* condition. As expected, all measures are correlated with ADM if values are drawn from a normal distribution(see Figure 4).

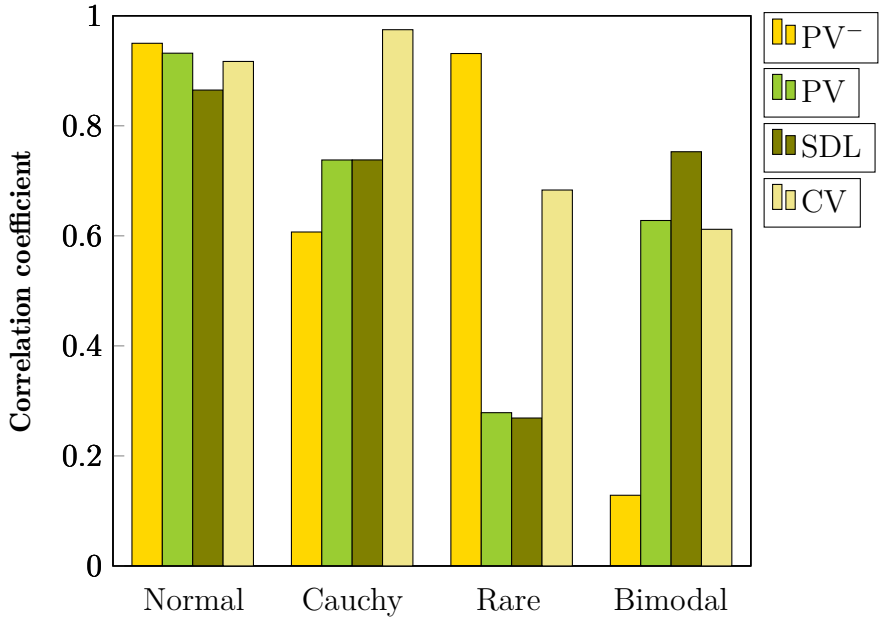


Figure 3: Correlation of PV^-, PV, SDL, and CV with average deviation from the mean (ADM), across 50.000 simulations of time series with 100 intervals. See caption of Figure 3 for details on the 4 conditions.

There is also a very high correlation between CV and ADM in the *Cauchy* condition, which shows that the pathological behavior of CV in this condition is due to its dependence on the mean. Similarly, in the *Rare* condition PV^- is strongly correlated with ADM, indicating that the average deviation is more important for the measure than the population crash. The correlation is lower for PV and SDL which are also more sensitive to the extent of the population crash. Conversely, in the *Bimodal* condition PV^- is only weakly correlated with ADM. This demonstrates that the index loses sensitivity for pairs of abundances if the differences between them are already large, but remains sensitive to changes that are often of less environmental concern. All four measures correlate with ADM for certain distributions, but the new measure PV has the advantage that the correlations with ADM for the three types of population dynamics that are not Gaussian are not very high.

Concluding Remarks

Population variability is an important concept in biological theory and practice, but standard measures have various shortcomings. They use deviations from the mean as the benchmark of variability which introduces a bias if the distribution of abundances is not Gaussian. Unlike previously described proportional measures the new measures quantify proportional variability relative to lower abundances, which makes them more sensitive if some abundances are very low relative to the other abundances in a time series. Given current environmental trends, this is a crucial property of a variability measure. It makes the proposed measures appropriate for the investigation of population dynamics from a conservation perspective. Other advantages are the flexible and unbiased evaluation of zero counts and an intuitive scale based on proportions. The new method to calculate them is computationally efficient and allows researchers to use these proportional measures to analyze the variability in longer time series.

Code availability

Jupyter Notebooks with code used for analyses are available at
`https://github.com/FT81/proportional-variability`

References

Fernández-Martínez, M., Vicca, S., Janssens, I. A., Carnicer, J., Martín-Vide, J., & Peñuelas, J. (2018). The consecutive disparity index, D: a measure of temporal variability in ecological studies. *Ecosphere, 9*(12), e02527.

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.

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.