OPEN ACCESS Freely available online

PLOS ONE

Quantifying Proportional Variability

Joel P. Heath*, Peter Borowski

Department of Mathematics, University of British Columbia, Vancouver, British Columbia, Canada

Abstract

Real quantities can undergo such a wide variety of dynamics that the mean is often a meaningless reference point for measuring variability. Despite their widespread application, techniques like the Coefficient of Variation are not truly proportional and exhibit pathological properties. The non-parametric measure Proportional Variability (PV) [1] resolves these issues and provides a robust way to summarize and compare variation in quantities exhibiting diverse dynamical behaviour. Instead of being based on deviation from an average value, variation is simply quantified by comparing the numbers to each other, requiring no assumptions about central tendency or underlying statistical distributions. While PV has been introduced before and has already been applied in various contexts to population dynamics, here we present a deeper analysis of this new measure, derive analytical expressions for the PV of several general distributions and present new comparisons with the Coefficient of Variation, demonstrating cases in which PV is the more favorable measure. We show that PV provides an easily interpretable approach for measuring and comparing variation that can be generally applied throughout the sciences, from contexts ranging from stock market stability to climate variation.

Citation: Heath JP, Borowski P (2013) Quantifying Proportional Variability. PLoS ONE 8(12): e84074. doi:10.1371/journal.pone.0084074

Editor: Shu-Dong Zhang, Queen's University Belfast, United Kingdom

Received May 30, 2013; **Accepted** November 20, 2013; **Published** December 30, 2013

Funding: NSERC Post Doctoral Fellowship to JPH and Supervisor funding to PB. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: heath.joel@gmail.com

Introduction

Understanding the variability of a quantity is a fundamental concept. The concept is generally considered intuitive, and techniques for measuring variability are rarely given a second thought, despite well established pathological issues [2–4]. Variation is typically based on calculating the average deviation from the mean. This of course assumes that the mean is a meaningful starting point for measuring variation. Given the central limit theorem, this may be a reasonable assumption in terms of statistical sampling; however, in terms of dynamics, real quantities rarely fluctuate about a central point and can exhibit a diverse spectrum of dynamics ranging from simple oscillation to chaos and noise [5]. There is nothing normal (Gaussian) about these dynamics and the mean can therefore be a misleading reference point for measuring variation. A common ground for measuring and comparing overall variation among quantities undergoing different dynamics requires a framework that is not secondarily based on a measure of central tendency.

There has been extensive confusion in several disciplines about the appropriate way to measure variation [2–4,6,7]. Given the standard deviation is related to the mean [8], the most commonly used and advocated technique for measuring variation on a proportional scale is the Coefficient of Variation (CV), computed as the standard deviation divided by the mean. Despite its widespread application, it has many pathological properties that can lead to inappropriate interpretation of results. In particular, it is not a truly proportional measure of variability, as it is not bounded by an upper value of 1. In contexts like population ecology, rare events are known to severely bias the CV, whereas a robust measure of variation should not rely on subjective decisions about what is rare and common nor involve inappropriately weighting or excluding data [9]. In the context of population dynamics, Heath [1] developed a simple solution to this issue: simply compare the numbers to each other rather than to an average. Using numerical simulations, it was shown that this technique is not biased by rare events or non-Gaussian dynamics, allows more accurate estimation of long term variability from short term data sets, and allows robust summarization and comparison of variability (or inversely, stability) among quantities undergoing very different dynamics. The present research uses mathematical proofs and an analytical approach to demonstrate desirable properties of Proportional Variability (PV), resolving important standing issues and providing a general replacement for the Coefficient of Variation. Variability is one of the most fundamental concepts in the sciences, and is particularly important for understanding contemporary issues including economic and environmental change. PV provides an intuitive and robust common ground for measuring and comparing variation on a proportional scale, and a new paradigm for concepts of variability.

Methods

Proportional Variability (PV) is based on a ratio comparison of all numbers. For a given data set of n non-negative points $z_i \geq 0$, there will be $C = n(n-1)/2$ unique pairwise combinations of (z_i, z_j), for which we calculate the relative difference $D(z_i, z_j)$. PV is therefore defined as:

$$\mathrm{PV} = \frac{1}{C}\sum_{\text{comb.}} D(z_i,z_j), \quad \text{where}$$
$$D(z_i,z_j) = \frac{|z_i - z_j|}{\max(z_i,z_j)} = 1 - \frac{\min(z_i,z_j)}{\max(z_i,z_j)}. \tag{1}$$

Unlike CV, the domain of PV is on a truly proportional scale [0,1]. Like the CV, this provides a summary of variation where the chronology of the data is irrelevant, and PV effectively considers variation at all scales or time lags. However, by ordering the data in a sequence of increasing magnitude, we can demonstrate some very desirable characteristics of a truly proportional measure of variability. Of course, if the data is constant, PV = 0, and the series is a simple horizontal line. If there is variability, the ordered series will be increasing, the steepness depending on the extent of variability. If the ordered series is linearly increasing, i.e., as an arithmetic sequence, and in the case the starting element of this sequence is zero, it is simple to prove that PV will equal exactly 0.5 independent of sample size n:

$$\mathrm{PV}_{\text{arithmetic}} = 1 - \frac{2}{n(n-1)}\sum_{j=2}^{n}\sum_{i=1}^{j-1}\frac{i}{j}$$
$$= 1 - \frac{1}{2}\frac{2}{n(n-1)}\left(\frac{n(n+1)}{2} - n\right) = \frac{1}{2}. \tag{2}$$

If variability as measured by PV is greater than 0.5, the values in the ordered series are increasing non-linearly. At the extreme, if variation increases exponentially with time as an ordered geometric series ($z_i = z_0 q^i$ with the common ratio $q > 1$), PV will approach a value of 1 as the sample size goes to infinity:

$$\mathrm{PV}_{\text{geometric}} = 1 - \frac{2}{n(n-1)}\sum_{j=2}^{n}\sum_{i=1}^{j-1}\frac{q^i}{q^j} = 1$$
$$- \frac{2}{n(n-1)}\frac{1}{1-q}\left(\frac{1-(1/q)^{n-1}}{q-1} - (n-1)\right) \overset{n\to\infty}{\to} 1 \tag{3}$$

for $q > 1$, and a similar proof is true for $q < 1$. This demonstrates very desirable and intuitive attributes of a proportional measure of variability. In contrast, CV has a domain that is theoretically $[0,\infty)$, rather inappropriate for a measure interpreted as being proportional.

If a system is normal (Gaussian), the Coefficient of Variation (and many other metrics based upon a measure of central tendency), is particularly excellent at providing useful first and second order summaries of the data (e.g. mean and standard deviation, respectively). By considering continuous probability distributions, we will demonstrate that PV behaves quantitatively similar to the CV under these conditions. This is desirable, but as we describe, the assumptions required for using CV will never be fully met by any systems of real (positive) quantities. We demonstrate that unlike CV, PV can be calculated for continuous distributions with an undefined mean. We conclude by considering bimodal distributions (e.g. simple oscillations), and show that PV is a cure for the pathological properties of the CV when otherwise stable populations exhibit rare events.

Results

For a continuous probability distribution $p(z)$ of a non-negative real variable $z \geq 0$, PV can be computed in the following way:

$$\mathrm{PV} = 1 - 2\int_0^\infty \int_{z_i}^\infty \frac{z_i}{z_j} p(z_i)p(z_j)\mathrm{d}z_j\mathrm{d}z_i. \tag{4}$$

In this section, we will analytically solve Eq. 4 for different continuous probability distributions, but first, we will consider the Gaussian or 'normal' distribution, where we solve Eq. 4 numerically. While rarely considered in practice, one issue with assuming data conform to a normal continuous distribution is the implication that some, albeit perhaps a small amount of numbers, can be negative. This is obviously an incorrect assumption for real positive quantities such as population abundances or temperature (the latter being analyzed in *Kelvins*). PV does not require this false assumption and is only appropriate for positive quantities. Therefore, to avoid contributions from negative values, we center the mean of the normal distributions at least two standard deviations in the positive domain, set the distribution equal to zero at negative values and renormalize the remaining part accordingly.

PV behaves quantitatively similar to CV across a wide range of distributions and qualitatively similar for very fat distributions (Figure 1, i.e., CV is larger for fat distributions as extreme values in the tails of the distribution are given substantially more weight due to deviations from the mean being squared in calculating the standard deviation). This indicates PV is a useful replacement for the CV under normal conditions. Because PV is independent of the mean, it also allows statistical analysis of variation vs. the mean, an inappropriate analysis for the CV.

As another test, we compute CV and PV for a uniform distribution centered at m with width w ($m > w/2$). Equation 4 leads to:

$$\mathrm{PV}_{\text{uniform}} = \frac{1}{w^2}\left[\left(m - \frac{w}{2}\right)^2 \ln\left(\frac{m + w/2}{m - w/2}\right) - mw + w^2\right] \quad \text{and}$$
$$\mathrm{CV}_{\text{uniform}} = \frac{w}{2\sqrt{3}m}. \tag{5}$$

These expressions for PV and CV are very close to each other for $w \ll m$, though they differ close to $w = 2m$ with $\mathrm{PV}_{\text{uniform}} = 0.5$ and $\mathrm{CV}_{\text{uniform}} = 1/\sqrt{3} \approx 0.577$. PV is therefore quantitatively similar to CV under normal conditions, providing an appropriate replacement. In a further example, we calculate CV and PV for the exponential distribution

$$p(z) = \lambda \exp(-\lambda z). \tag{6}$$

After some algebra, it turns out that both $\mathrm{CV}_{\text{exp}} = 1$ and $\mathrm{PV}_{\text{exp}} = 2(1 - \ln(2)) \approx 0.6137$, are independent of the decay constant λ in this case.

PV has the advantage compared to CV and other mean-based metrics, that it can be computed even when a mean is not defined. The Pareto distribution (e.g. [10]) is useful for describing many observable phenomenon such as the skewed distribution of wealth. It is defined as:

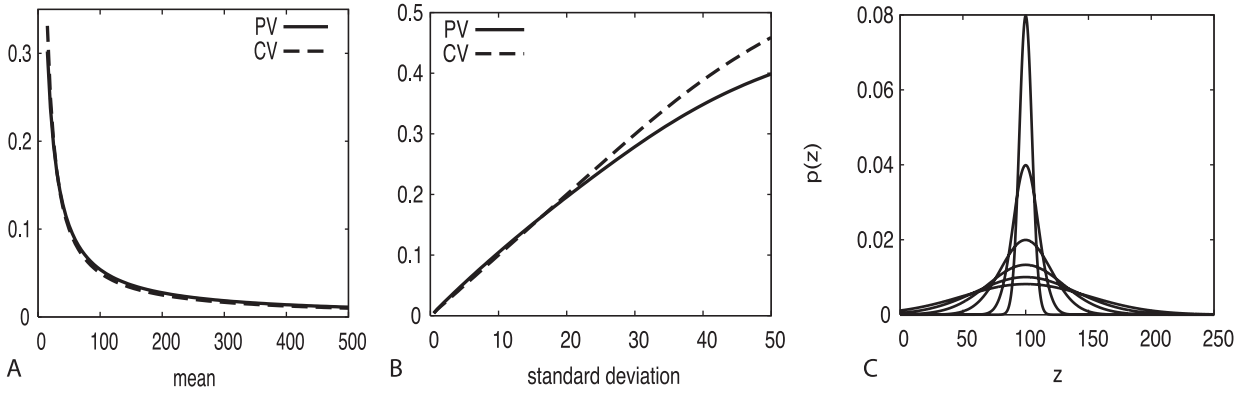

Figure 1. Proportional Variability (PV) and the Coefficient of Variation (CV) correspond closely and quantitatively over a large range of parameters for the Gaussian distribution. Before truncating negative numbers and renormalizing, (A) standard deviation = 5, and (B) stable mean = 100 with increasing standard deviation (as per C, $\sigma = 5,10,20,30,40,50$). Both CV and PV have been obtained by numerically solving the defining integral equations.
doi:10.1371/journal.pone.0084074.g001

$$p(z) = \frac{k}{z_{\min}}\left(\frac{z_{\min}}{z}\right)^{k+1} \quad \text{for } z \geq z_{\min};$$
$$p(z) = 0 \quad \text{for } z < z_{\min} \tag{7}$$

with the two parameters scale $z_{\min} > 0$ and shape $k > 0$. For the Pareto distribution, a mean is only defined if the shape parameter k is larger than one, and a second moment (including CV) is only defined for $k > 2$. PV is a first order measure of variability and can be computed using Equation 4 for all k to $\mathrm{PV}_{\mathrm{Pareto}} = 1/(k+1)$, independent of the scale parameter $z_{\min}$.

The mean is also not defined for the heavy-tailed Cauchy-Lorentz (CL) distribution (e.g. [10])

$$p(z) = \frac{1}{\pi}\frac{s}{s^2 + (z-t)^2} \tag{8}$$

which is described by a scale s and location parameter t. We restrict our analysis to positive CL distributions, considering the positive half of the distribution with a location parameter $t = 0$ (and an according normalization factor of two). The CV cannot be calculated for the CL distribution. Interestingly, for this special case $(t = 0)$, the definite integrals in Eq. 4 can be solved analytically, leading to the result $\mathrm{PV}_{\mathrm{CL}} = 2/3$, which is independent of the scale parameter s. This indicates that contributions near $z = 0$ and large z perfectly balance each other out as the scale parameter changes for this distribution. Heath [1] demonstrated numerically that PV is much more appropriate when rare events occur, which is the case for heavy-tailed Cauchy-Lorentz distributions. This is because by comparing each number to every other number, rare events are evaluated in direct relation to their frequency (i.e., how rare they are). PV therefore solves the problem of rare events and allows calculating variability without inappropriate decisions to include outliers or not.

Bimodal distributions offer another intuitive way to compare measures of variability for rare events or other mixed distributions such as those produced by simple oscillatory dynamics. As a clear demonstration, we analyse discrete distributions with N_A and $N_B = n - N_A$ counts at only two magnitudes, z_A and $z_B = z_A + \zeta$. In this case, with $\eta = N_B/(N_A + N_B)$

$$\mathrm{PV}_{\mathrm{bimodal}} = \frac{\zeta}{z_A + \zeta}\frac{2n\eta(1-\eta)}{n-1} \quad \text{and}$$
$$\mathrm{CV}_{\mathrm{bimodal}} = \frac{\zeta}{z_A + \zeta\eta}\sqrt{\eta(1-\eta)}. \tag{9}$$

The first notable observation here is that when exchanging N_A and N_B (i.e., replacing η with $1-\eta$ while keeping sample size n constant), PV behaves symmetrically, maintaining the same value, whereas CV changes in magnitude. The difference between PV and CV (and indeed other central tendency based measures) becomes very obvious when considering large deviations ζ between the two magnitudes (z_A and z_B) and large differences between the two counts (i.e., η close to zero or to one). The latter is the regime of rare events. In Figure 2, we chose a particularly demonstrative example, a perfectly stable population with all elements at the same common value and only a single rare event. Increasing the sample size n while maintaining a single rare event ($N_B = 1$) means that the proportion of rare events will be decreasing monotonically due to increasing counts of the common value. As seen in Figure 2, PV decreases monotonically with sample size as expected and is appropriate, however CV actually increases initially for low sample sizes before beginning a slow decrease. This characteristic of CV is particularly pathological, and an inappropriate description of variability for this quantity which is otherwise always stable.

A well-known measure of distribution that behaves similarly to PV for the case of a bimodal distribution is the relative mean difference (RMD) or related Gini coefficient (defined as half of RMD – e.g. [11]). For the remaining distributions discussed in the present report, expressions for RMD can be found elsewhere (e.g. [12]). RMD (and therefore Gini coefficient) is based on the mean, which is why RMD can not be computed for mean-less distributions like Cauchy-Lorentz or Pareto with $k \leq 1$.

Discussion

This treatment demonstrates the wide applicability and robustness of Proportional Variability (PV). While CV is only appropriate for normal distributions, PV behaves the same as CV under these conditions, but also functions as desired for all other possible forms of dynamics. This makes PV particularly appealing

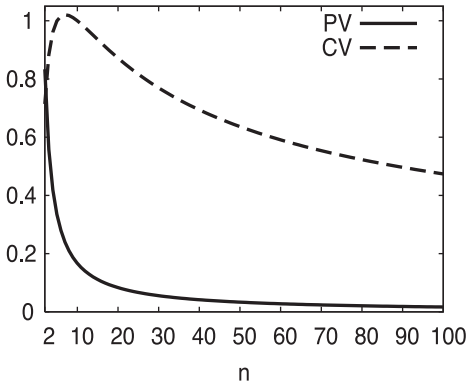

Figure 2. For a quantity which is stable at all time intervals except for a single rare event, CV displays a pathological increase in magnitude with an increase in sample size n before slowly decreasing. PV appropriately declines monotonically to zero as the frequency of the rare event decreases with increasing sample size for this otherwise stable quantity. In this example $z_A = 10$, $\zeta = 50$, $N_B = 1$.
doi:10.1371/journal.pone.0084074.g002

for summarizing and comparing variability in data that can undergo a wide variety of dynamics, such as in non-linear economic, physical and biological systems [5]. Heath [1] also numerically demonstrated that PV behaves the same as the spectral exponent (a useful gold standard) when used to evaluate more time more variation (reddened spectrum), whereas CV inappropriately suggests spectral reddening in stationary time series. Furthermore, PV allows substantially more accurate estimates of known long term variability from short term sampling, for a variety of distributions [1]. Recent work supports the finding that PV is robust to rare events, showing lower standard error in PV compared to other metrics using jackknife estimates [13]. PV has now been applied to quantify variability in populations of a variety of species [14,15], in oceanography [16], characterizing frequency distributions [17], and quantifying variability in backscatter measurements from acoustic surveys [18]. The present treatment provides an analytical basis for the further development of PV, and many additional characteristics and applications are likely to be discovered as it is implemented in various contexts by researchers with different expertise. Overall, our results indicate a strong case for the general adoption of PV as a standard measure of variability: it is a truly proportional first order measure of variability, avoids issues associated with standard techniques, and provides a robust common ground to summarize and compare variability in systems undergoing a wide variety of dynamic behaviour. We advocate PV as a useful common ground for evaluating and comparing stability (1/PV) and variation throughout the sciences. It is our intention to encourage critical thinking about philosophies of variability, and to raise skepticism and caution in applying and interpreting other approaches such as the Coefficient of Variation. Redefining our paradigm of variability will be particularly important for addressing contemporary issues of economic and climatic variability, and for establishing relationships between variability in physical, biological and socioeconomic quantities.

Acknowledgments

We are grateful to many individuals for their thoughtful comments on this approach as it has developed, particularly to members of the Mathematical Biology program in the Department of Mathematics at the University of British Columbia for feedback on the present treatment.

Author Contributions

Conceived and designed the experiments: JPH PB. Performed the experiments: PB. Analyzed the data: PB JPH. Contributed reagents/materials/analysis tools: PB JPH. Wrote the paper: JPH PB.

References

1. Heath J (2006) Quantifying temporal variability in population abundances. Oikos 115: 573–581185.
2. Pimm S (1991) The Balance of Nature?: Ecological Issues in the Conservation of Species and Communities. University of Chicago Press.
3. McArdle B, Gaston K, Lawton J (1990) Variation in the size of animal populations: patterns, problems and artefacts. The Journal of Animal Ecology 59: 439–454.
4. Gaston KJ, McArdle BH (1994) The temporal variability of animal abundances: measures, methods and patterns. Philosophical Transactions of the Royal Society of London Series B: Biological Sciences 345: 335–358.
5. May RM (1976) Simple mathematical models with very complicated dynamics. Nature 261: 459–467.
6. Martin JD, Gray LN (1971) Measurement of relative variation: Sociological examples. American Sociological Review 36: 496–502.
7. Kokko H, Mackenzie A, Reynolds JD, Lindström J, Sutherland WJ (1999) Measures of inequality are not equal. The American Naturalist 154: 358–382.
8. Taylor L (1961) Aggregation, variance and the mean. Nature 189: 732–735.
9. Halley JM (1996) Ecology, evolution and 1/f-noise. Trends in Ecology & Evolution 11: 33–37.
10. Johnson NL, Kotz S, Balakrishnan N (1994) Continuous Univariate Distributions, Vol. 1 (Wiley Series in Probability and Statistics). Wiley-Interscience, 2nd edition.
11. Sen A (1973) On economic inequality. Oxford University Press.
12. Wikipedia (2013). Mean difference – Wikipedia, the free encyclopedia. [Online; accessed 14-October-2013].
13. Dochtermann NA, Peacock MM (2010) Differences in population size variability among populations and species of the family salmonidae. Journal of Animal Ecology 79: 888–896.
14. Corbineau A, Rouyer T, Cazelles B, Fromentin JM, Fonteneau A, et al. (2008) Time series analysi s of tuna and swordfish catches and climate variability in the indian ocean (1968–2003). Aquatic Living Resources 21: 277–285.
15. Lamb RJ, MacKay PA, Alyokhin A (2011) Population variability and persistence of three aphid pests of potatoes over 60 years. Canadian Entomologist 143: 91–101.
16. Rouyer T, Fromentin JM, Stenseth NC (2010) Environmental noise affects the fluctuations of atlantic large pelagics. Progress in Oceanography 86: 267–275.
17. Jovani R, Mavor R (2011) Group size versus individual group size frequency distributions: a nontrivial distinction. Animal Behaviour 82: 1027–1036.
18. Adam P (2009) Quantifying spatial and temporal variability of distribution patterns in acoustic backscatter density with synoptic airborne LIDAR. M.sc. thesis, University of Washington.