

A variability index for spatial series of rates

Indici di Variabilità per Serie Territoriali di Rapporti Statistici

Maria Rita Sebastiani

Università “La Sapienza” - Dipartimento di Studi Geoeconomici, Linguistici, Statistici,
Storici per l'Analisi Regionale, via del Castro Laurenziano 9, 00161, Roma,
mariarita.sebastiani@uniroma1.it

Riassunto: Lo scopo del lavoro è quello di introdurre un indice normalizzato di variabilità per studiare l'entità della variazione presente in una serie di rapporti statistici. Dopo aver definito un indice assoluto opportuno, è necessario determinare gli estremi del campo di variazione di questo. Lo studio delle distribuzioni che massimizzano o minimizzano gli indici di variabilità è stato già trattato in letteratura, nel caso delle sole distribuzioni statistiche secondo variabili non negative. Per le serie di rapporti statistici non è altrettanto facile individuare la serie massimizzante, soprattutto se i tassi assumono per natura soltanto valori compresi tra 0 e 1. Si considerano tassi costruiti con variabili non negative di natura trasferibile. Determinate le condizioni analitiche per l'esistenza del minimo e del massimo, è stato implementato un algoritmo di ricerca in R.

Keywords: normalized index, variability.

1. Introduction

Statistical rates are useful tools for studying the spatial dynamics of phenomena affecting populations. Given a spatial series of rates, it might be important to outline the general situation regarding the entire set of sites, by measuring the variation of the rates within the series. For this purpose, it is useful to adopt a normalized variability index, that is expressed as a number independent of the general average size or the unit of measurement, with a range between 0 and 1. By using such an index, one can compare the variability of two or more spatial series. Moreover, by referring to the bounded range, we can evaluate how high the degree of variation is within a series.

In the literature, there are many statistical indices of variability. In general, given a univariate distribution, to obtain a normalized index from an absolute index, we need to determinate the bounds of the absolute index. The statistical distribution with minimum or maximum variability has been discussed widely in the literature. Specifically for the variance, solutions have been found when the variable is non-negative. These solutions cannot be applied *tout court* to series of rates. The main problem derives from the fact that, if the rates considered are defined to comply with some logical and/or formal constraints, the same constraints must hold within the series with extreme variability. Moreover, some rates can assume by nature negative values.

The aim of this paper is to develop and to test methodologies for introducing a new normalized index of variability, analogous to the normalized variance, that can be used for spatial series of rates. Here we consider rates where the variable in the numerator is non-negative and transferable by nature. The case of rates that can assume both negative and non-negative values is more complex and requires further investigation.

2. The absolute index

Let $S = \{1, \dots, j, \dots, k\}$ be a set of label representing k sites, where $k \in \mathcal{N}$ is fixed. Let x_j , p_j and $t_j = x_j / p_j$ be respectively the variable, the population and the rate observed for the j -th site. More generally, we can think of p_j as a positive term for normalizing x_j . Let a and p be respectively the total of the x_j values and the p_j over S . Let $\bar{t}$ be the weighted average rate of the t_j over S : $\bar{t} = \left(\sum_{j=1}^k t_j p_j \right) / \sum_{j=1}^k p_j = a / p$.

Let $\mathcal{X}$ be the set of the possible series $(X_1, \dots, X_j, \dots, X_k)$, corresponding to $(p_1, \dots, p_j, \dots, p_k)$, such that each has total equal to a and average rate equal to $\bar{t}$: $\mathcal{X} = \left\{ (X_1, \dots, X_j, \dots, X_k) : \sum_{j=1}^k X_j = a \right\} = \left\{ (X_1, \dots, X_j, \dots, X_k) : \bar{t} = a / p \right\}$.

Here, we assume that all the X_j are non-negative and transferable.

If the purpose is to measure the variability of the $T_j = X_j / p_j$ with regard to the average size of the phenomenon, it is convenient to express the index as the r -th moment of $(T_j - \bar{t})$. Since $\bar{t}$ is a weighted average, we believe that it is right to define the index by weighting the differences $(T_j - \bar{t})^r$ by p_j . Taking $r = 2$, we obtain:

$$^2S_t^2 = \frac{1}{p} \sum_{j=1}^k (T_j - \bar{t})^2 p_j = \frac{1}{p} \sum_{j=1}^k \left(\frac{X_j}{p_j} - \frac{a}{p} \right)^2 p_j. \quad (1)$$

3. Normalizing the index: the minimum and the maximum values

If the index (1) varies in a finite range with known extremes, we can define the corresponding normalized index by the formula:

$$0 \leq ^2S_{t,norm}^2 = \frac{^2S_t^2 - \min}{\max - \min} \leq 1. \quad (2)$$

3.1 Determining the minimum value

For all series $(X_1, \dots, X_j, \dots, X_k) \in \mathcal{X}$, we have by definition that: $^2S_t^2 \geq 0$. Specifically, the series that achieves the minimum of 0 for (1) is such that:

$$^2S_t^2 = 0 \Leftrightarrow T_j = X_j / p_j = a / p \Leftrightarrow X_j = a \cdot p_j / p \text{ for every } j = 1, \dots, k. \quad (3)$$

3.2.1 Determining the maximum value: the general situation

Given the hypotheses $0 \leq X_j \leq a$ for each $j = 1, \dots, k$, we have:

$$X_i \leq a \Rightarrow \frac{X_i}{p_i} \leq \frac{a}{p_i} \Rightarrow \left(\frac{X_i}{p_i} - \frac{a}{p} \right)^2 \leq \left(\frac{a}{p_i} - \frac{a}{p} \right)^2 \quad i = 1, \dots, k. \quad (4)$$

Thus, the value of (1) can be increased by putting $X_i = a$ in one term of the sum. Since $\sum_{j=1}^k X_j = a$, we must have that $X_j = 0$ for each $j \neq i$. In symbols, the maximizing series

has the form $\left(\underbrace{0, \dots, 0}_{i-1}, a, \underbrace{0, \dots, 0}_{k-i} \right)$ and all the rates are null apart from that for the i -th

unit, that is equal to $t_i^* = a / p_i$. By applying definition (1) to this series, we obtain:

$$\max {}^2S_i^2 = \bar{t} \cdot (t_i^* - \bar{t}) = \bar{t} \cdot (a / p_i - \bar{t}). \quad (5)$$

In general, there may be several maximizing series each corresponding to a specific choice of the i -th unit. Amongst them, since (5) depends inversely on p_i , the absolute maximum for ${}^2S_i^2$ is achieved when the only non-null term corresponds to the minimum population, that is taking: $i = \operatorname{argmin}_{1 \leq j \leq k} \{p_1, \dots, p_j, \dots, p_k\}$ in (5).

3.2.2 Determining the maximum: the case of rates varying between 0 and 1

Consider the case where the terms X_j are part of the corresponding p_j ($X_j \leq p_j$), so that the rates vary between 0 and 1. Here the minimizing series remains the one given in Section 3.1. The maximizing series, in addition to satisfying the conditions given in Section 3.2.1, must comply with the constraints $X_j \leq p_j$; thus, this series has the same form as given in Section 3.2.1, but here we also have that: $i = \operatorname{argmin} \{p_j \text{ such that } p_j \geq a\}$.

The requirement stated between parentheses is not itself sufficient for determining the maximizing series. Indeed, it may be the case that $p_j < a$ for every $j = 1, \dots, k$ in which case we need to consider all the series where the total a is distributed among two or more units of S . Substituting $X_k = a - \sum_{j=1}^{k-1} X_j$ and $p_k = p - \sum_{j=1}^{k-1} p_j$ into (1), we obtain:

$${}^2S_i^2(X_1, \dots, X_{k-1}) = \sum_{j=1}^{k-1} \frac{X_j^2}{p p_j} + \frac{\left(a - \sum_{j=1}^{k-1} X_j \right)^2}{p \left(p - \sum_{j=1}^{k-1} p_j \right)} - \frac{a^2}{p^2}. \quad (6)$$

Given the hypotheses about the X_j , the variability increases as a becomes more concentrated among the k sites. Specifically, the variability is a maximum if a is distributed between two units, for instance i and l provided that $a = X_i + X_l \leq p_i + p_l$. Setting: $X_j = 0$ for each $j \neq i, l$ and $X_l = a - X_i$, where X_i is fixed and $0 < X_i \leq p_i$, we see that (6) represents the equation of a concave parabola. If, on the other hand, it is the case that $p_i + p_l < a$ for all pairs i, l , we have that the variability is maximized if a is distributed among three units, for example i , l and m and if $a = X_i + X_l + X_m \leq p_i + p_l + p_m$. By suitable substitutions, analogous to those given above, (6) becomes the equation of a surface turned upside down. Generalizing, suppose that for each subset $I_n = \{i_1, i_2, \dots, i_n\} \subset S$ containing n units ($n \leq k-1$), it holds that $\sum_{j \in I_n} p_j < a$. Then, the variability is maximized if a can be distributed among $n+1$ units; without loss in generality, let us indicate these units by the set $I_{n+1} = \{h_1, h_2, \dots, h_n, h_{n+1}\} \subset S$. This means that (1) can be increased by setting $X_j = 0$ for each $j \notin I_{n+1}$ and $X_{h_{n+1}} = a - (X_{h_1} + X_{h_2} + \dots + X_{h_n})$, where each X_{h_j} in the parentheses is fixed and $0 < X_{h_j} \leq p_{h_j}$. This solution is acceptable only if $\sum_{j \in I_{n+1}} p_j \geq a$. Suppose that this condition does not hold for every proper n -dimensional subset of S . Then, the $(X_1, \dots, X_{k-1})$ that maximizes (6) is such that a is distributed amongst all the k units of S . By varying $(X_1, \dots, X_{k-1})$, equation (6) represents a solid figure. We can show that all the elements of the hessian matrix $\mathbf{H}$ of (6) are positive and that both the determinant of $\mathbf{H}$ and that of the matrix $\mathbf{H}_{k-1, k-1}$, obtained by dropping the last row and the last column of $\mathbf{H}$, are also positive. Consequently, the solid represented by (6) has the shape of a bell turned upside down with the maximum being on the border. If we substitute in (6) $X_1 = \tilde{x}_1$, $X_2 = \tilde{x}_2$, ..., $X_{k-2} = \tilde{x}_{k-2}$ where $1 \leq \tilde{x}_j \leq p_j$ for each $j = 1, 2, \dots, k-2$, which represents the intersection of the solid (6) with corresponding planes, we obtain the equation of a concave parabola with the maximum being on the border ($X_{k-1} = 1$ or $X_{k-1} = p_{k-1}$). By varying $(\tilde{x}_1, \tilde{x}_2, \dots, \tilde{x}_{k-2})$ we have different parabolas. The absolute maximum of (6) corresponds to the parabola with maximum gradient. The gradient is an increasing function of $(\tilde{x}_1, \tilde{x}_2, \dots, \tilde{x}_{k-2})$. Thus, for the series $(p_1, \dots, p_j, \dots, p_k)$ there are two possible solutions. For determining the maximum numerically, we implemented an algorithm in R. Since assigning a label $1, \dots, k$ to each site is arbitrary, the algorithm works on all possible $k!$ label permutations.

References

- Benedetti C. (1981). *Istituzioni di Statistica*, Veschi, Roma.
- Brizzi M. (1996). *Le misure relative di variabilità. Lineamenti metodologici e applicazioni*, Esculapio, Bologna.
- Naddeo A. (1984). *Statistica di base*, Edizioni Kappa, Roma.
- Venables W.N., Smith D.M. and the R Development Core Team (2002). *An Introduction to R*, version 1.6.1.