

Conditioning on the Mean in the Evaluation of Dispersion of Ordinal or Natural-value Variables

Dispersione condizionata alla Media su Variabili ordinali o naturali

Andrea Scagni

Dipartimento di Statistica e Matematica applicata - Università di Torino

Abstract

In molti casi le risposte ai sondaggi d'opinione vengono fornite su scale ordinali, relativamente ad un livello di interesse, soddisfazione o simili. In indagini su larga scala diviene spesso necessario sintetizzare i risultati con un indice di locazione e scala, sulla base della codifica delle risposte su numeri naturali da 1 a k , anche se approcci più complessi cercano di definire metriche "ottimali". La valutazione della dispersione mediante la varianza o indici specifici (D di Leti) può essere allora normalizzata dato k o il parametro di locazione, da cui è strettamente dipendente. Il comportamento degli indici condizionati viene studiato e comparato per la deviazione standard e il D di Leti.

Keywords: ordinal variables, Leti's D index, dispersion, mean conditioning

1. Introduction

Responses to opinion surveys are frequently given on ordinal scales, where the individual has to specify a level of interest, agreement, satisfaction, or similar. In complex and large-scale surveys it is then often useful to summarize the results, replacing frequency distributions with some indexes showing location and dispersion of the responses. An example is the annual *Student's opinion survey on university teaching* in Turin, where 16 items regarding the quality of lectures are investigated on about 3200 courses, with results delivered to some 2000 professors.

In these cases the responses are often encoded with natural numbers from 1 through k , even though more complex approaches to define an optimal metric have been developed (Gifi, 1990; Delvecchio, 2004).

The evaluation of dispersion based on the ranked responses can then be performed using the standard deviation σ , and the behaviour of the statistic can be shown to be practically equivalent to that of statistics specifically devised for categorical ordinal variables, as Leti's D (Leti, 1983). Both σ and D share the same range (asymptotically for D), and their values, although not directly linked, are very close for all distributions for a wide range of k (Scagni, 2005). Also, the decomposition property *within* and *between* groups holds for both statistics (Grilli, Rampichini, 2002).

In these cases dispersion statistics can thus be normalized, since their maximum is a function of k . However it makes little sense to compare "normalized" dispersion on very differently located distributions, as their value is strictly dependent on the location of the distribution: when responses are concentrated near 1 or k their potential dispersion is more limited than when they are concentrated near $k/2$.

In this paper the use of normalized dispersion statistics conditioned on a location parameter is proposed. Such statistics are derived from both standard deviation and Leti's D ; their behaviour is studied and compared.

2. Natural-value variables mean and dispersion

Let A be an ordinal categorical variable, with k response categories $a_1, a_2, \dots, a_k$. Let X be the discrete variable derived from A through the rank transformation $a_i \rightarrow x_i = i; i = 1, 2, \dots, k$. If n_i denotes the associated absolute frequency and $N = \sum_{i=1}^k n_i; f_i = n_i/N; F_i = \sum_{j=1}^i f_j$, the mean and variance of X are:

$$\bar{x} = \sum_{i=1}^k i (F_i - F_{i-1}) = \sum_{i=1}^k i F_i - \sum_{i=1}^{k-1} (i+1) F_i = k F_k + \sum_{i=1}^{k-1} i F_i - \sum_{i=1}^{k-1} (i+1) F_i = k - \sum_{i=1}^{k-1} F_i$$

$$\sigma^2 = \sum_{i=1}^k (i - \bar{x})^2 f_i = \sum_{i=1}^k i^2 f_i - \bar{x}^2 = \sum_{i=1}^k i^2 (F_i - F_{i-1}) - \left(k - \sum_{i=1}^{k-1} F_i \right)^2$$

In this case both statistics can be written as functions of the cumulative distribution function only. The same holds of course for Leti's dispersion index for ordinal variables $D = 2 \sum_{i=1}^{k-1} F_i (1 - F_i)$. Both σ and D can vary from 0 through $(k-1)/2$ (if $n \rightarrow \infty$ for D with even N), since if $f_i = 1/2; i = 1, k; f_i = 0; i = 2, 3, \dots, k-1$ then:

$$\sigma_{max}^2 = (k^2 + 1)/2 - [(k+1)/2]^2 = (k^2 - 2k + 1)/4 = ((k-1)/2)^2 \Rightarrow \sigma_{max} = (k-1)/2$$

It is then possible, at least in principle, to define the corresponding normalized statistic $\sigma_r = \sigma/\sigma_{max} = \sigma/[(k-1)/2]$ and $D_r = D/[(k-1)/2]$.

For a comparison of the behaviour of σ and D , see Scagni (2004).

Such "marginal" normalization, however, ignores that the range of σ and D varies sharply if considered conditionally on a fixed location parameter, and can actually reach $(k-1)/2$ only when $\bar{x} = (k+1)/2$. A way to avoid this is to normalize dispersion statistics *conditionally on the observed mean*, as shown in the following paragraphs.

3. Maximum conditional dispersion: characterization of the distribution

Lemma 1. Given $1 \leq \bar{x} \leq k$, consider the class of discrete natural-valued variables $\{X\}$ on $(1; k)$ with mean $\bar{x}$. The maximum of σ_x^2 on such class is reached if and only if:

$$f_1 = (\bar{x} - k)/(1 - k); f_k = (1 - \bar{x})/(1 - k); f_j = 0 \quad \forall j = 2, \dots, k-1$$

Proof. Given $\bar{x}$, σ_x^2 is a monotonous increasing function of the second moment:

$$M(X^2) = \sum_{i=1}^k i^2 (F_i - F_{i-1}) = k^2 + \sum_{i=1}^{k-1} i^2 F_i - \sum_{i=1}^{k-1} (i+1)^2 F_i = k^2 + \sum_{i=1}^{k-1} [i^2 - (i+1)^2] F_i = k^2 + \sum_{i=1}^{k-1} (-2i - 1) F_i = k^2 - \sum_{i=1}^{k-1} (1 + 2i) F_i = k^2 - 3 F_1 - \sum_{i=2}^{k-1} (1 + 2i) F_i$$

Since $\sum_{i=1}^{k-1} F_i = k - \bar{x} \Rightarrow F_1 = k - \bar{x} - \sum_{i=2}^{k-1} F_i$, this is also:

$$M(X^2) = k^2 - 3 F_1 - \sum_{i=2}^{k-1} (1 + 2i) F_i = k^2 - 3(k - \bar{x}) + 2 \sum_{i=2}^{k-1} (1 - i) F_i \quad (1)$$

Since $2 \sum_{i=2}^{k-1} (1 - i) F_i < 0$, and $F_i \geq F_1$, the maximum of (1) given $\bar{x}$ is reached when $F_i = F_1$, but this implies that the distribution is concentrated on values 1 and k only. Then the mean becomes $\bar{x} = k - \sum_{i=1}^{k-1} F_i = k - (k-1) F_1 = k - (k-1) f_1$; solving for f_1 ,

$f_1 = (\bar{x} - k)/(1 - k)$ is obtained and therefore $f_k = 1 - (\bar{x} - k)/(1 - k) = (1 - \bar{x})/(1 - k)$. ■

Lemma 2. Given $1 \leq \bar{x} \leq k$, consider the class of discrete natural-valued variables $\{X\}$ on $(1; k)$ with mean $\bar{x}$. The maximum of D on such class is reached if and only if:

$$f_1 = (\bar{x} - k)/(1 - k); \quad f_k = (1 - \bar{x})/(1 - k); \quad f_j = 0 \quad \forall j = 2, \dots, k-1$$

Proof. Conditional on $\bar{x}$, D can be written as $D/2 = [(k - \bar{x}) - \sum_{i=1}^{k-1} F_i^2]$ (since $\sum_{i=1}^{k-1} F_i = k - \bar{x}$), so that the minimum of $\sum_{i=1}^{k-1} F_i^2$ must be found with respect to the $k-1$ unknowns F_i subject to the constraint $\sum_{i=1}^{k-1} F_i = k - \bar{x} = c$. The Lagrangian function is $g(F_i; \lambda) = \sum_{i=1}^{k-1} F_i^2 - \lambda(\sum_{i=1}^{k-1} F_i - k + \bar{x})$, and its partial derivatives are:

$$\partial g(F_i; \lambda) / \partial F_i = 2F_i - \lambda; \quad \partial g(F_i; \lambda) / \partial \lambda = \sum_{i=1}^{k-1} F_i - k + \bar{x} \quad (2)$$

Setting both derivatives in (2) to zero, each F_i , $i=1, \dots, k-1$, must be equal to the constant $\lambda/2$. The values of λ is thus given by:

$$\sum_{i=1}^{k-1} F_i = \sum_{i=1}^{k-1} \frac{\lambda}{2} = k - \bar{x} \Rightarrow (k-1) \frac{\lambda}{2} = k - \bar{x} \Rightarrow \lambda = 2 \frac{k - \bar{x}}{k-1} = 2 \frac{\bar{x} - k}{1 - k} \quad (3)$$

Again, $F_i = \lambda/2$, $i=1, \dots, k-1$ implies that the distribution is concentrated on values 1 and k only, and by (3) $f_1 = (\bar{x} - k)/(1 - k)$ and $f_k = (1 - \bar{x})/(1 - k)$. ■

On a natural-value variable X , therefore, the maximum of both σ and D - given any value of $\bar{x}$ - is attained when the distribution is concentrated on its extremes 1 and k .

4. Maximum conditional dispersion: normalized statistics

It is now straightforward to derive the maximum of both σ and D conditional on $\bar{x}$:

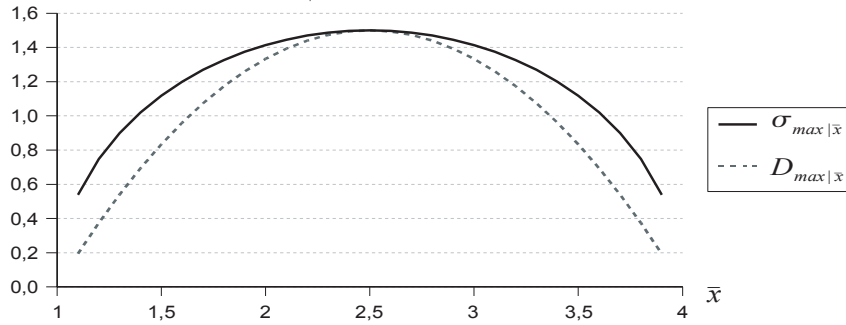
$$\begin{aligned} \sigma_{\max|\bar{x}}^2 &= \sum_{i=1}^k x_i^2 f_i - \bar{x}^2 = \frac{\bar{x} - k}{1 - k} + k^2 \frac{1 - \bar{x}}{1 - k} - \bar{x}^2 = \bar{x} \left(\frac{1 - k^2}{1 - k} \right) - \bar{x}^2 - k = \\ &= \bar{x}(1 + k) - \bar{x}^2 - k = (k - \bar{x})(\bar{x} - 1) \Rightarrow \sigma_{\max|\bar{x}} = [(k - \bar{x})(\bar{x} - 1)]^{1/2} \end{aligned}$$

$$\begin{aligned} D_{\max|\bar{x}} &= 2 \sum_{i=1}^{k-1} F_i(1 - F_i) = 2(k-1) \frac{\bar{x} - k}{1 - k} \cdot \frac{1 - \bar{x}}{1 - k} = -2 \frac{(\bar{x} - k)(1 - \bar{x})}{1 - k} = \\ &= 2 \frac{\bar{x} - \bar{x}^2 - k + \bar{x}k}{k-1} = 2 \frac{\bar{x}(1 + k - \bar{x}) - k}{k-1} = 2 \frac{(k - \bar{x})(\bar{x} - 1)}{k-1} \end{aligned}$$

and define the associated *conditional normalized statistics* $\sigma_{r|\bar{x}} = \sigma / [(k - \bar{x})(\bar{x} - 1)]^{1/2}$ and $D_{r|\bar{x}} = (k-1)D / [2(k - \bar{x})(\bar{x} - 1)]$. Both $\sigma_{\max|\bar{x}}$ and $D_{\max|\bar{x}}$ reduce to their unconditional maximum $(k-1)/2$ when $\bar{x} = (k+1)/2$; both are functions of the same core $(k - \bar{x})(\bar{x} - 1)$, but $D_{\max|\bar{x}}$ is also inversely linked to the range $(k-1)$. In other words, the maximum dispersion (conditional on $\bar{x}$) is a quadratic function of how $\bar{x}$ itself partitions the range of X : more uneven partitions allow for less dispersion.

Figure 1 shows the similar behaviour of $\sigma_{\max|\bar{x}}$ and $D_{\max|\bar{x}}$ as functions of $\bar{x}$, even though the latter has a somewhat wider range.

Figure 1: Values of $\sigma_{max|\bar{x}}$ and $D_{max|\bar{x}}$ with $k=4$ as functions of $\bar{x}$

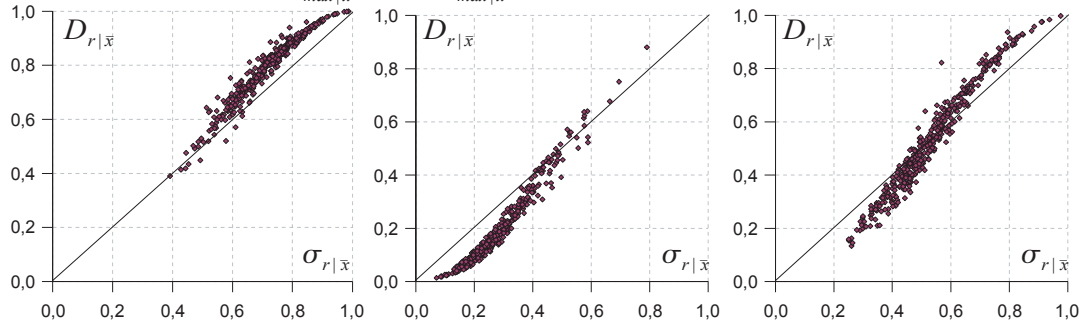


When considered as functions of k for any fixed $\bar{x}$, $\sigma_{max|\bar{x}}$ and $D_{max|\bar{x}}$ show some asymptotic differences: while the first is obviously unbounded, the latter is finite:

$$\lim_{k \rightarrow \infty} D_{max|\bar{x}} = 2 \lim_{k \rightarrow \infty} \frac{(k - \bar{x})(\bar{x} - 1)}{k - 1} = 2(\bar{x} - 1)$$

Empirical behaviour of $\sigma_{r|\bar{x}}$ and $D_{r|\bar{x}}$ is remarkably similar, even though they are not one-to-one functions, nor is one of them more discriminating than the other. Simulation studies show that $\sigma_{r|\bar{x}} > D_{r|\bar{x}}$ for low values, while $D_{r|\bar{x}} > \sigma_{r|\bar{x}}$ for high values. Figure 2 shows the results of 3 sets of 400 ($k=4$) replications with high ($f_i=0.25; 0.25; 0.25; 0.25$), low ($f_i=0.02; 0.06; 0.91; 0.01$) and intermediate ($f_i=0.08; 0.12; 0.68; 0.12$) dispersion.

Figure 2. Simulated $\sigma_{max|\bar{x}}$ and $D_{max|\bar{x}}$ on 3 different $k=4$ random distributions



References

- Leti G. (1983) *Statistica descrittiva*, Il Mulino, Bologna, 290-297.
- Delvecchio F. e G. (2004) *La quantificazione mediante algoritmi genetici come strumento per il confronto di giudizi medi espressi su scala ordinale*, Convegno "Strategie metodologiche per lo studio della transizione Università-Lavoro", Cleup, Padova, 199-220.
- Gifi, A. (1990) *Nonlinear multivariate analysis*, Chichester, UK, John Wiley.
- Grilli L., Rampichini C. (2002) Scomposizione della dispersione per variabili ordinali, *Statistica*, anno LXII, 1: 111-116.
- Scagni A. (2004) *Le opinioni degli studenti della didattica in università: sono attendibili le differenze nei risultati aggregati di facoltà?* Convegno "Outcomes 2004: Modelli di analisi della transizione università-lavoro", Cleup, Padova, 1, 23-38.