Chapter 1

MEASURING DISPERSION IN THE CONTEXT OF ORDERED QUALITATIVE SCALES

José Luis García-Lapresta[1] and Luis Borge[2]

[1] *PRESAD Research Group, BORDA Research Unit, IMUVA, Departamento de Economía Aplicada, Universidad de Valladolid, Spain*
lapresta@eco.uva.es

[2] *Departamento de Economía Aplicada, Universidad de Valladolid, Spain*
borge@eco.uva.es

Abstract In this contribution, we introduce a family of dispersion measures in the context of ordered qualitative scales.

Keywords: Qualitative scales; Dispersion; Ordinal proximity measures.

1. Introduction

The main purpose of this contribution is to measure the dispersion associated with the different results of a variable when they form a set of linguistic terms obtained after examining different qualities of people, services, etc, ordered, but with a non-homogeneous and non-quantifiable distance between the linguistic terms.

There is in the literature a great variety of dispersion measures associated with a data set, being the most used the standard deviation. For a very long time, the normality assumption was present in most of the statistical studies and its estimator in combination with the sample mean provide excellent efficiency from the practical side due to its precision to adjust the data and from the capacity of giving rise to theoretical results in the sample distributions. However, most of the estimators of dispersion presents in literature do not hold

for our purposes, because they are based on numerical values obtained in the sample.

When the data that we want to study are obtained from a statistical behaviour different from the normal one, there are atypical observations and losses of efficiency on the classical estimators. The growth of statistical methods and their increasing use in different areas of research poses the study of alternative estimators which are less sensible to deviations to the normal. In the middle of the last century there was an important revival of the robust statistics. The concept of robustness is related to the fact that a estimator is less sensible to small changes in the statistical hypothesis about the models or about the presence of atypical observations, being the breakdown point one of the ways to measure the robustness of a estimator which is related to the percentage of polluted observations of a sample.

The most well known measure of localization of the robust statistical is the median which is estimated ordering the values obtained in a sample, while the sample median is obtained as the central value of the ordered data, if the number of data is odd, and as the average of the central data, if it is even. When the data of a sample is like the ones we are interested in, we may use this localization measurement when the number of data n is odd. When we are dealing with the robust dispersion the most used measurement is the interquartile range. This method is not the best one because a estimator is more robust if it is obtained as a median of the absolute deviation of the median of the sample, also known as MAD. If we have an ordered sample of a variable $(x_{(1)},\ldots,x_{(n)})$, its median, $\text{med}_i(x_i)$ is defined

$$\text{MAD} = \text{med}_j|x_j - \text{med}_i(x_i)|.$$

To obtain this estimator, it is necessary to evaluate the median two times. The first one between all the linguistic terms of the data. The second median, after ordering the different proximities between the linguistic terms, we choose the median of these proximities.

Rousseeuw and Croux [8] presented two alternative robust estimators to the MAD and, as the interquartile range, they do not require to center the data to respect to the sample median. The first of them is

$$S_n = \text{med}_i\left\{\text{med}_j|x_i - x_j|\right\},$$

which for any x_i value we obtain the median of the differences of the form $\{|x_i - x_j|, j = 1,\ldots,n\}$, which gives us n medians, being the estimator S_n the median of this set of medians.

The second estimator proposed by Rousseeuw and Croux, which they call Q_n, is based only on the differences between data values. If $\{x_{(1)},\ldots,x_{(n)}\}$ is the ordered sample, let $\mathscr{D} = \{x_{(i)} - x_{(j)}, i > j\}$ be the set of interpoint distances in which the number of elements is $m = \binom{n}{2}$. Then, the estimate is

defined as the k-th order statistic of $\mathscr{D}$, where $k = \binom{[n/2]+1}{2}$ and that is roughly half the number of observations. These estimator has attractive properties (see Maronna, et al.**)

**Some absolute dispersion measures used in Statistics are the range, the variance, the mean deviation, the standard deviation, the absolute Gini index, etc. (see Martínez-Panero et al. [7]).

The rest of the contribution is organized as follows. Section 2 is devoted to introduce the notation and some basic notions that are necessary for defining the two families of dispersion measures we propose in the setting of ordered qualitative scales. Section 3 present these two families of dispersion measures and includes some illustrative examples. Section 4 contains some properties. Finally, Section 5 concludes with some remarks.

2. Notation and Basic Notions

First we present the notion of ordinal proximity measure.

2.1 Ordinal Proximity Measures

Let $\mathscr{L} = \{l_1, \ldots, l_g\}$ be an ordered qualitative scale arranged from the lowest to the highest linguistic terms, $l_1 < l_2 < \cdots < l_g$, with $g \geq 3$.

In order to recall the notion of ordinal proximity measure on $\mathscr{L}$, introduced by García-Lapresta and Pérez-Román [4], we shall use a linear order $\Delta = \{\delta_1, \ldots, \delta_h\}$, with $\delta_1 \succ \cdots \succ \delta_h$, for representing different degrees of proximity among the terms of $\mathscr{L}$, being δ_1 and δ_h the maximum and minimum degrees, respectively.

It is important noticing that the elements of Δ have no meaning and they only represent different degrees of proximity.

As usual in the setting of linear orders, $\delta_r \succeq \delta_s$ means $\delta_r \succ \delta_s$ or $\delta_r = \delta_s$; $\delta_r \prec \delta_s$ means $\delta_s \succ \delta_r$; and $\delta_r \preceq \delta_s$ means $\delta_r \prec \delta_s$ or $\delta_r = \delta_s$.

Given a weak order $\trianglelefteq$ on $\mathscr{L}^n$, with $\triangleleft$ we denote the asymmetric part of $\trianglelefteq$, i.e., $\boldsymbol{x} \triangleleft \boldsymbol{y} \Leftrightarrow$ not $\boldsymbol{y} \trianglelefteq \boldsymbol{x}$.

DEFINITION 1.1 *([4]) An* ordinal proximity measure on $\mathscr{L}$ with values in Δ *is a mapping* $\pi : \mathscr{L}^2 \longrightarrow \Delta$, *where* $\pi(l_r, l_s) = \pi_{rs}$ *means the degree of proximity between* l_r *and* l_s, *satisfying the following conditions:*

1 Exhaustiveness*: For every* $\delta \in \Delta$, *there exist* $l_r, l_s \in \mathscr{L}$ *such that* $\delta = \pi_{rs}$.

2 Symmetry*:* $\pi_{sr} = \pi_{rs}$, *for all* $r, s \in \{1, \ldots, g\}$.

3 Maximum proximity*:* $\pi_{rs} = \delta_1 \Leftrightarrow r = s$, *for all* $r, s \in \{1, \ldots, g\}$.

4 Monotonicity*:* $\pi_{rs} \succ \pi_{rt}$ *and* $\pi_{st} \succ \pi_{rt}$, *for all* $r, s, t \in \{1, \ldots, g\}$ *such that* $r < s < t$.

The previous conditions are independent (see García-Lapresta and Pérez-Román [4, Prop. 1]).

Every ordinal proximity measure $\pi : \mathscr{L}^2 \longrightarrow \Delta$ can be represented by a $g \times g$ symmetric matrix with coefficients in Δ, where the elements in the main diagonal are $\pi_{rr} = \delta_1$, for every $r = 1, \ldots, g$:

$$\begin{pmatrix} \pi_{11} & \cdots & \pi_{1s} & \cdots & \pi_{1g} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ \pi_{r1} & \cdots & \pi_{rs} & \cdots & \pi_{rg} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ \pi_{g1} & \cdots & \pi_{gs} & \cdots & \pi_{gg} \end{pmatrix} = (\pi_{rs}).$$

This matrix is called the *proximity matrix associated with* π.

Taking into account the conditions appearing in Definition 1.1, it is only necessary to show the upper half proximity matrix

$$\begin{pmatrix} \delta_1 & \pi_{12} & \pi_{13} & \cdots & \pi_{1(g-1)} & \pi_{1g} \\ & \delta_1 & \pi_{23} & \cdots & \pi_{2(g-1)} & \pi_{2g} \\ & & & \cdots & \cdots & \cdots \\ & & & & \delta_1 & \pi_{(g-1)g} \\ & & & & & \delta_1 \end{pmatrix}.$$

As shown in García-Lapresta and Pérez-Román [4, Prop. 2], the minimum proximity between linguistic terms is only reached when comparing the extreme linguistic terms: $\pi_{rs} = \delta_h \Leftrightarrow (r,s) \in \{(1,g),(g,1)\}$.

The cardinality of Δ is located between the cardinality of $\mathscr{L}$ and a polynomial of degree 2 of that cardinality (see García-Lapresta and Pérez-Román [4, Prop. 4]):

$$g \leq h \leq \frac{g \cdot (g-1)}{2} + 1.$$

2.2 Medians

Following García-Lapresta and Pérez-Román [5], we now introduce the median operator in the setting of ordinal degrees of proximity.

Given a vector of ordinal degrees of proximity $\boldsymbol{\delta} = (\delta_1, \ldots, \delta_p) \in \Delta^p$, we arrange its components in a decreasing fashion, from the highest to the lowest degrees. If p is odd, then the median of $\boldsymbol{\delta}$ is unique, say $\delta_r \in \Delta$. However, if p is even, then $\boldsymbol{\delta}$ has two medians, say $\delta_s, \delta_t \in \Delta$ such that $s \leq t$, i.e., $\delta_s \succeq \delta_t$. In order to unify the assignment of medians, we consider the pair of medians (δ_r, δ_r) and (δ_s, δ_t) whenever p is odd and even, respectively.

More formally, given the *set of feasible medians* $\Delta_2 = \{(\delta_r, \delta_s) \in \Delta^2 \mid r \leq s\}$, the *median operator* is the mapping

$$M : \bigcup_{p=1}^{\infty} \Delta^p \longrightarrow \Delta_2$$

that assigns the corresponding pair of medians to each vector of ordinal degrees.

For ordering the pairs of medians of ordinal proximities, consider the linear order $\succeq_2$ on Δ_2 defined as

$$(\delta_r, \delta_s) \succeq_2 (\delta_t, \delta_u) \Leftrightarrow \begin{cases} r+s < t+u \\ \text{or} \\ r+s = t+u \text{ and } s-r \leq u-t, \end{cases} \tag{1}$$

for all $(\delta_r, \delta_s), (\delta_t, \delta_u) \in \Delta_2$.

It is easy to see that if $r+s = t+u$, then $s-r \leq u-t \Leftrightarrow r \geq t \Leftrightarrow s \leq u$.

3. Dispersion Measures

In this section we present two families of dispersion measures in the setting of ordered qualitative scales equipped with ordinal proximity measures. The first family generalizes the most basic dispersion measure, the range.

3.1 Range-based dispersion measures

Given $n \geq 2$, let $D_R : \mathscr{L}^n \longrightarrow \Delta$ be the mapping defined as

$$D_R(\boldsymbol{x}) = \pi(\min \boldsymbol{x}, \max \boldsymbol{x}), \tag{2}$$

for every $\boldsymbol{x} \in \mathscr{L}^n$.

Based on the linear order $\succeq$ on Δ, we introduce the weak order $\trianglelefteq_R$ on $\mathscr{L}^n$ defined as

$$\boldsymbol{x} \trianglelefteq_R \boldsymbol{y} \Leftrightarrow D_R(\boldsymbol{x}) \succeq D_R(\boldsymbol{y}),$$

with the meaning of the dispersion in $\boldsymbol{x}$ is lower than or equal to in $\boldsymbol{y}$ (with respect to D_R).

EXAMPLE 1.2 Consider the ordered qualitative scale $\mathscr{L} = \{l_1, l_2, l_3, l_4\}$ and the vectors $\boldsymbol{x} = (l_1, l_2, l_2, l_3)$, $\boldsymbol{y} = (l_3, l_3, l_4, l_4) \in \mathscr{L}^4$. We want to compare the dispersion in these vectors with respect to three different ordinal proximity measures.

1 If $\mathscr{L}$ is equipped with the ordinal proximity measure

$$\pi : \mathscr{L}^2 \longrightarrow \Delta = \{\delta_1, \ldots, \delta_7\}$$

with associated proximity matrix

$$\begin{pmatrix} \delta_1 & \delta_2 & \delta_4 & \delta_7 \\ & \delta_1 & \delta_3 & \delta_6 \\ & & \delta_1 & \delta_5 \\ & & & \delta_1 \end{pmatrix},$$

we have $D_R(\boldsymbol{x}) = \pi_{13} = \delta_4 \succ \delta_5 = \pi_{34} = D_R(\boldsymbol{y})$. Thus, $\boldsymbol{x} \lhd_R \boldsymbol{y}$.

2 If $\mathscr{L}$ is equipped with the ordinal proximity measure

$$\pi : \mathscr{L}^2 \longrightarrow \Delta = \{\delta_1, \ldots, \delta_7\}$$

with associated proximity matrix

$$\begin{pmatrix} \delta_1 & \delta_4 & \delta_6 & \delta_7 \\ & \delta_1 & \delta_3 & \delta_5 \\ & & \delta_1 & \delta_2 \\ & & & \delta_1 \end{pmatrix},$$

we have $D_R(\boldsymbol{x}) = \pi_{13} = \delta_6 \prec \delta_2 = \pi_{34} = D_R(\boldsymbol{y})$. Thus, $\boldsymbol{y} \lhd_R \boldsymbol{x}$.

3 If $\mathscr{L}$ is equipped with the ordinal proximity measure

$$\pi : \mathscr{L}^2 \longrightarrow \Delta = \{\delta_1, \ldots, \delta_4\}$$

with associated proximity matrix

$$\begin{pmatrix} \delta_1 & \delta_2 & \delta_3 & \delta_4 \\ & \delta_1 & \delta_2 & \delta_3 \\ & & \delta_1 & \delta_2 \\ & & & \delta_1 \end{pmatrix},$$

we have $D_R(\boldsymbol{x}) = \pi_{13} = \delta_3 \prec \delta_2 = \pi_{34} = D_R(\boldsymbol{y})$. Thus, $\boldsymbol{y} \lhd_R \boldsymbol{x}$.

3.2 Gini-based dispersion measures

We now introduce a new family of dispersion measures in the mentioned framework. It is based on the Gini index ([6]) and it is closely related to the scale estimator appearing in Shamos [9, p. 260] and Bickel and Lehmann [3, p. 38] in the setting of real numbers (see Rousseeuw and Croux [8, p. 1277]).

Given $n \geq 2$, let $D_G : \mathscr{L}^n \longrightarrow \Delta_2$ be the mapping defined as

$$D_G(\boldsymbol{x}) = M(\pi(x_i, x_j)_{i<j}), \tag{3}$$

for every $\boldsymbol{x} = (x_1, \ldots, x_n) \in \mathscr{L}^n$.

Based on the linear order $\succeq_2$ on Δ_2 defined in (1), we introduce the weak order $\trianglelefteq_G$ on $\mathscr{L}^n$ defined as

$$\boldsymbol{x} \trianglelefteq_G \boldsymbol{y} \Leftrightarrow D_G(\boldsymbol{x}) \succeq_2 D_G(\boldsymbol{y}),$$

with the meaning of the dispersion in $\boldsymbol{x}$ is lower than or equal to in $\boldsymbol{y}$ (with respect to D_G).

Since some vectors can share the same pair of medians, it is necessary to devise a tie-breaking process for ordering the vectors. We propose to use a sequential procedure based on Balinski and Laraki [1] (see Balinski and Laraki [2] for practical examples). It consists of withdrawing the pair of medians of the vectors that are in a tie, and then selecting the new pairs of medians of the remaining proximity degrees for the corresponding vectors. The process continues until the ties are broken. It is important to note that different vectors never are in a final tie.

EXAMPLE 1.3 Consider Example 1.2 and the same three ordinal proximity measures.

1 We have

$$D_G(\boldsymbol{x}) = M(\pi_{12}, \pi_{12}, \pi_{13}, \pi_{22}, \pi_{23}, \pi_{23}) = M(\delta_2, \delta_2, \delta_4, \delta_1, \delta_3, \delta_3) = (\delta_2, \delta_3)$$

and

$$D_G(\boldsymbol{y}) = M(\pi_{33}, \pi_{34}, \pi_{34}, \pi_{34}, \pi_{34}, \pi_{44}) = M(\delta_1, \delta_5, \delta_5, \delta_5, \delta_5, \delta_1) = (\delta_5, \delta_5).$$

Since $(\delta_2, \delta_3) \succ_2 (\delta_5, \delta_5)$, we have $\boldsymbol{x} \triangleleft_G \boldsymbol{y}$.

2 We have

$$D_G(\boldsymbol{x}) = M(\pi_{12}, \pi_{12}, \pi_{13}, \pi_{22}, \pi_{23}, \pi_{23}) = M(\delta_4, \delta_4, \delta_6, \delta_1, \delta_3, \delta_3) = (\delta_3, \delta_4)$$

and

$$D_G(\boldsymbol{y}) = M(\pi_{33}, \pi_{34}, \pi_{34}, \pi_{34}, \pi_{34}, \pi_{44}) = M(\delta_1, \delta_2, \delta_2, \delta_2, \delta_2, \delta_1) = (\delta_2, \delta_2).$$

Since $(\delta_3, \delta_4) \prec_2 (\delta_2, \delta_2)$, we have $\boldsymbol{y} \triangleleft_G \boldsymbol{x}$.

3 We have

$$D_G(\boldsymbol{x}) = M(\pi_{12}, \pi_{12}, \pi_{13}, \pi_{22}, \pi_{23}, \pi_{23}) = M(\delta_2, \delta_2, \delta_3, \delta_1, \delta_2, \delta_2) = (\delta_2, \delta_2)$$

and

$$D_G(\boldsymbol{y}) = M(\pi_{33}, \pi_{34}, \pi_{34}, \pi_{34}, \pi_{34}, \pi_{44}) = M(\delta_1, \delta_2, \delta_2, \delta_2, \delta_2, \delta_1) = (\delta_2, \delta_2).$$

Consequently, in $\boldsymbol{x}$ and $\boldsymbol{y}$ the dispersion is the same. If we apply the tie-breaking procedure, then we have $D_G(\boldsymbol{x}) = M(\delta_1, \delta_2, \delta_2, \delta_3) = (\delta_2, \delta_2)$ and $D_G(\boldsymbol{y}) = M(\delta_1, \delta_1, \delta_2, \delta_2) = (\delta_1, \delta_2)$. Since $(\delta_2, \delta_2) \prec_2 (\delta_1, \delta_2)$, we finally have $\boldsymbol{y} \triangleleft_G \boldsymbol{x}$.

4. Properties

Let $\mathscr{L}$ be an ordered qualitative scale equipped with an ordinal proximity measure $\pi : \mathscr{L}^2 \longrightarrow \Delta$. We say that π is *totally uniform* if $\pi_{r(r+t)} = \pi_{s(s+t)}$ for all $r,s,t \in \{1,\ldots,g-1\}$ such that $r+t \leq g$ and $s+t \leq g$.

Let $N : \mathscr{L} \longrightarrow \mathscr{L}$ be the *negation operator* defined as $N(l_r) = l_{g+1-r}$, for every $r \in \{1,\ldots,g\}$.

Given $k \in \{1-g,\ldots,g-1\}$, let $T_k : \mathscr{L} \longrightarrow \mathscr{L}$ be the *translation operator* defined as $T_k(l_r) = l_{r+k}$, for every $r \in \{1,\ldots,g\}$ such that $r+k \leq g$.

In the following proposition we establish some properties of the mappings introduced in (2) and (3). They are related to the ones considered in Martínez-Panero et al. [7] in a quantitative context.

PROPOSITION 1.1 *Consider the mappings $D_R : \mathscr{L}^n \longrightarrow \Delta$ and $D_G : \mathscr{L}^n \longrightarrow \Delta_2$ defined in (2) and (3), respectively, and its extensions $\widetilde{D}_R : \bigcup_{n=2}^{\infty} \mathscr{L}^n \longrightarrow \Delta$ and $\widetilde{D}_G : \bigcup_{n=2}^{\infty} \mathscr{L}^n \longrightarrow \Delta_2$. The following properties hold:*

1. Symmetry*: $D_R(x_{\sigma(1)},\ldots,x_{\sigma(n)}) = D_R(x_1,\ldots,x_n)$ and $D_G(x_{\sigma(1)},\ldots,x_{\sigma(n)}) = D_G(x_1,\ldots,x_n)$, for every permutation $\sigma : \{1,\ldots,n\} \longrightarrow \{1,\ldots,n\}$ and every $(x_1,\ldots,x_n) \in \mathscr{L}^n$.*
2. Invariance for replications*: $\widetilde{D}_R(\overbrace{\boldsymbol{x},\ldots,\boldsymbol{x}}^{m}) = D_R(\boldsymbol{x})$ and $\widetilde{D}_G(\overbrace{\boldsymbol{x},\ldots,\boldsymbol{x}}^{m}) = D_G(\boldsymbol{x})$, for every $\boldsymbol{x} \in \mathscr{L}^n$ and any number $m \in \mathbb{N}$ of replications of $\boldsymbol{x}$.*
3. Minimum dispersion*: $D_R(x_1,\ldots,x_n) = \delta_1 \Leftrightarrow x_1 = \cdots = x_n$; and $D_G(l_r,\ldots,l_r) = (\delta_1,\delta_1)$, for every $l_r \in \mathscr{L}$.*
4. Anti-self-duality*: if π is totally uniform, then $D_R(N(x_1),\ldots,N(x_n)) = D_R(x_1,\ldots,x_n)$ and $D_G(N(x_1),\ldots,N(x_n)) = D_G(x_1,\ldots,x_n)$, for every $(x_1,\ldots,x_n) \in \mathscr{L}^n$.*
5. Invariance for translations*: if π is totally uniform, then $D_R(T_k(x_1),\ldots,T_k(x_n)) = D_R(x_1,\ldots,x_n)$ and $D_G(T_k(x_1),\ldots,T_k(x_n)) = D_G(x_1,\ldots,x_n)$, for every $(x_1,\ldots,x_n) \in \mathscr{L}^n$ and every $k \in \{1,\ldots,g-1\}$ such that $(T_k(x_1),\ldots,T_k(x_n)) \in \mathscr{L}^n$.*

5. Concluding Remarks

Acknowledgments

This contribution is dedicated to the memory of Pedro Gil. The first author gratefully acknowledges the funding support of the Spanish *Ministerio de Economía y Competitividad* (project ECO2016-77900-P) and ERDF.

References

[1] Balinski, M., Laraki, R. (2007). A theory of measuring, electing and ranking. *Proceedings of the National Academy of Sciences of the United States of America* 104, pp. 8720–8725.

[2] Balinski, M., Laraki, R. (2013). How best to rank wines: Majority Judgment. In: *Wine Economics: Quantitative Studies and Empirical Observations*, pp. 149–172, Palgrave-MacMillan.

[3] Bickel, P.J., Lehmann, E.L. (1979). Descriptive Statistics for nonparametric models IV: Spread. In: J. Jurečková (ed.) *Contributions to Statistics, Hájek Memorial Volume*, pp. 33–40, Academia, Prague.

[4] García-Lapresta, J.L., Pérez-Román, D. (2015). Ordinal proximity measures in the context of unbalanced qualitative scales and some applications to consensus and clustering. *Applied Soft Computing* 35, pp. 864–872.

[5] García-Lapresta, J.L., Pérez-Román, D. (2017). A consensus reaching process in the context of non-uniform ordered qualitative scales. *Fuzzy Optimization and Decision Making*. Forthcoming, DOI: 10.1007/s10700-016-9256-6.

[6] Gini, C. (1912). *Variabilità e Mutabilità*, Tipografia di Paolo Cuppini, Bologna.

[7] Martínez-Panero, M., García-Lapresta, J.L., Meneses, L.C. (2016). Multidistances and dispersion measures. In: T. Calvo Sánchez, J. Torrens Sastre (eds.) *Fuzzy Logic and Information Fusion*, Studies in Fuzziness and Soft Computing, pp. 123–134, Springer.

[8] Rousseeuw, P.J., Croux, C. (1993). Alternatives to the median absolute deviation. *Journal of the American Statistical Association* 88, pp. 1273–1283.

[9] Shamos, M.I. (1976). Geometry and Statistics: Problems at the interface. In: J.F. Traub (ed.) *New Directions and Recent Results in Algorithms and Complexity*, pp. 251–280, New York Academic Press.