DE GRUYTER
OPEN

STUDIES IN LOGIC, GRAMMAR
AND RHETORIC 50 (63) 2017

DOI: 10.1515/slgr-2017-0017

Grzegorz Lissowski
University of Warsaw

TWO MEASURES OF THE DEPENDENCE OF PREFERENTIAL RANKINGS ON CATEGORICAL VARIABLES[1]

Abstract. The aim of this paper is to apply a general methodology for constructing statistical methods, which is based on decision theory, to give a statistical description of preferential rankings, with a focus on the rankings' dependence on categorical variables. In the paper, I use functions of description errors that are based on the Kemeny and Hamming distances between preferential orderings, but the proposed methodology can also be applied to other methods of estimating description errors.

Keywords: decision theory, individual preferential ordering (ranking), group (social) preferential ordering (ranking), Kemeny distance, Hamming distance, generalized regression, measure of the intensity of a dependence, computational complexity.

1. Introduction

Information about how people rank the elements of a set with respect to a given criterion, e.g. occupations with respect to prestige or public institutions with respect to trust, tells us a great deal about their attitudes. However, many researchers are unfamiliar with methods of statistical description of this kind of data and, as a result, what is usually collected are not orderings of the whole set, but evaluations of its particular members on a scale of ordered categories. Many years ago, in response to requests from researchers, I proposed to address this by using the methods of aggregating individual preferences into social preferences developed by social choice theorists (Lissowski 1974a). But some features that are desirable from the perspective of social choice are no longer so when methods of statistical description are concerned. So I later proposed a new class of methods, based on the idea that one can treat statistical description as a solution to a decision problem. On this approach, parameters of various properties of

ISBN 978-83-7431-527-2 ISSN 0860-150X

statistical distribution are linked to one another by estimates of description errors.

This article presents some methods of measuring the intensity of statistical dependence of preferential rankings on other variables. A method (or rather a class of methods) to measure this kind of dependence will be proposed. The proposed method has been developed according to the methodology, mentioned above, of treating statistical description as a decision problem (Lissowski 1974b, 1976, 1977). More specifically, the intensity of statistical dependence is identified with the pragmatic value of information about preferential orderings provided by the categorical variable. The measure uses functions of description errors associated with distances between preferential orderings: the Kemeny distance and the Hamming distance. However, the methodology can also be applied to other methods of error estimation and other measures of distance between preferential orderings.

2. Individual preferential rankings

Let $A = \{a_1, \ldots, a_j, \ldots, a_m\}$ or $A = \{a, b, c, d, \ldots\}$ be a finite m-element *set of objects*, such as occupations, public institutions, values, motives or situations. This set of objects is being evaluated by an individual or a group of individuals with respect to a particular property or criterion.

An ordering of a set of objects can be represented as *a preferential ranking or ordering*, i.e. a sequence of the set's elements listed in decreasing order of assigned value, where elements that are assigned the same value are joined with a hyphen, e.g.

$$\mathbf{R}_1 : \quad a \;\; b \;\; c - d \qquad\qquad\qquad \mathbf{R}_2 : \quad b - c \;\; d \;\; a$$

This is an abbreviated notation for an order relation on set $\mathbf{A}$, i.e. a binary relation $\mathbf{R}$ satisfying the conditions of reflexivity, transitivity, and completeness. Formally, an order relation is a subset of the Cartesian product $\mathbf{A} \times \mathbf{A}$, or a set of ordered pairs of set $\mathbf{A}$ that stand to one another in relation $\mathbf{R}$ ($\mathbf{R} \subseteq \mathbf{A} \times \mathbf{A}$). For example,

$$\mathbf{R}_1 = \{(a,a),(a,b),(a,c),(a,d),(b,b),(b,c),(b,d),(c,c),(c,d),(d,d),(d,c)\}$$

$$\mathbf{R}_2 = \{(b,b),(b,c),(b,d),(b,a),(c,c),(c,b),(c,d),(c,a),(d,d),(d,a),(a,a)\}$$

By $a\mathbf{R}b$ we mean that element a is taken to be at least as valuable as element b. The preference relation, $\mathbf{R}$, can be divided into two distinct parts: the asymmetric relation of strict preference $\mathbf{P} = \{(a, b) : a\mathbf{R}b$ and it is not the case that $b\mathbf{R}a\}$ and the symmetric relation of indifference $\mathbf{I} = \{(a, b) : a\mathbf{R}b$ and $b\mathbf{R}a\}$. Therefore $\mathbf{R} = \mathbf{P} \cup \mathbf{I}$ and, in consequence, $|\mathbf{R}| = |P \cup \mathbf{I}| = |\mathbf{P}| + |\mathbf{I}|$, where $|\mathbf{B}|$ denotes the cardinality of set $\mathbf{B}$.

Relation $\mathbf{R}$ can be displayed as a matrix, wherein 1 stands for the relation obtaining between a pair of elements and 0 stands for the relation failing to obtain. This will be useful and convenient in what follows. The two relations $\mathbf{R}_1$ and $\mathbf{R}_2$ above can be represented as follows:

$$\mathbf{R}_1 = \begin{array}{c} \\ a \\ b \\ c \\ d \end{array} \begin{array}{c} \begin{array}{cccc} a & b & c & d \end{array} \\ \begin{bmatrix} 1 & 1 & 1 & 1 \\ 0 & 1 & 1 & 1 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 1 & 1 \end{bmatrix} \end{array} \qquad \mathbf{R}_2 = \begin{array}{c} \\ a \\ b \\ c \\ d \end{array} \begin{array}{c} \begin{array}{cccc} a & b & c & d \end{array} \\ \begin{bmatrix} 1 & 0 & 0 & 0 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 1 & 0 & 0 & 1 \end{bmatrix} \end{array}$$

Let $\mathbf{G} = \{1, 2, \ldots, h, \ldots, n\}$ denote a group of n individuals. Then by *the profile of preference rankings or orderings* of the set of n individuals we shall mean an n-tuple of individual preference rankings

$$\mathbf{\Pi}_G = \langle \mathbf{R}_1, \mathbf{R}_2 \ldots, \mathbf{R}_h, \ldots, \mathbf{R}_n \rangle,$$

where $\mathbf{R}_h$ is a preference ranking of person h.

3. Group preferential ranking

It is no easy task to determine a group preferential ranking given a profile of individual preferences (composed of the various preferences of the members of group $\mathbf{G}$). The problem of aggregating individual preferences into a group preference has long been investigated by social choice theory. The purpose of this normative theory is to establish a group preference that will enable us to compare various possible solutions from the social point of view or allow the group to make the best (i.e., most democratic and fairest) decisions. The aim of this article is to identify the group preference that best reflects the underlying individual preferential orderings. The two aims, the normative and the descriptive, are related, but not identical (see Lissowski 2000).

The group preference that best reflects the set of underlying individual preferential rankings should be selected in such a way as to minimize the expected value of loss associated with the description, where the expected value of loss is a function of description errors. Statistical description is thus treated as an optimal solution to a certain decision problem. This is how a whole unified class of methods of statistical description is constructed on the basis of decision theory.

There are many ways in which to estimate how well a group preference reflects the set of underlying individual preferential rankings. The simplest and probably most important ones consist in comparing the preference relations between pairs of objects in the individual preferences with those in the group preference. Table 1 shows the types of possible errors involved.

Table 1

Possible description errors associated with a group preference relation of group G vis-à-vis an individual preference relation of person *h* between a pair of objects

Individual preference (an element of the described set)	Group preference (a potential description)		
	$a\mathbf{P}_G b$	$a\mathbf{I}_G b$	$b\mathbf{P}_G a$
$a\mathbf{P}_h b$	e_1	e_2	e_3
$a\mathbf{I}_h b$	e_4	e_5	e_6
$b\mathbf{P}_h a$	e_7	e_8	e_9

In three cases, namely e_1, e_5, and e_9, there is no description error, since the individual preference relation between the two objects is the same as the group preference. In the remaining cases, individual preference diverges from group preference, but the description errors involved differ in character and magnitude. Yet one can find similarities between some of them.

Errors e_3 and e_7 are of a similar type: a strict individual preference is the reverse of a strict group preference. If the objects are treated in a neutral manner, the two types of error should be evaluated as much the same. Errors e_4 and e_6 are also similar. In both, the individual preference is an indifference, whereas the group preference is a strict preference. They should be evaluated as the same for much the same reason (of neutrality). A comparison of errors e_2 and e_8 is more complicated. In both cases, the group preference is an indifference, whereas the individual preferences are strict (opposing) preferences. Assuming neutrality toward the objects, the errors should be on a par, but the group members may evaluate them differently from errors e_4 and e_6.

Below, I present the simplest two functions for estimating description errors; they are called loss functions in the literature. I proposed to use them in the statistical description of the set of preferential orderings in a paper published in 1974. The first function, ℓ_1, takes the value 2 if a group preference is the opposite of an individual preference, the value 1 if one of the preferences is a strict preference while the other is an indifference, and the value 0 when both preferences are the same, so that there is no description error. The second function, ℓ_2, takes 1 when the description of a group preference does not reflect an individual preference (irrespective of what kind of error is involved), and 0 when the description is correct.

$$\ell_1[e(a,b)] = \begin{cases} 0 & \text{if } e \in \{e_1, e_5, e_9\} \\ 1 & \text{if } e \in \{e_2, e_4, e_6, e_8\} \\ 2 & \text{if } e \in \{e_3, e_7\} \end{cases}$$

$$\ell_2[e(a,b)] = \begin{cases} 0 & \text{if } e \in \{e_1, e_5, e_9\} \\ 1 & \text{if } e \in \{e_2, e_3, e_4, e_6, e_7, e_8\} \end{cases}$$

Table 2

The first function of description errors $\ell_1[e(a,b)]$

Individual preference (an element of the described set)	Group preference (a potential description)		
	$a\mathbf{P}_G b$	$a\mathbf{I}_G b$	$b\mathbf{P}_G a$
$a\mathbf{P}_h b$	0	1	2
$a\mathbf{I}_h b$	1	0	1
$b\mathbf{P}_h a$	2	1	0

Table 3

The second function of description errors $\ell_2[e(a,b)]$

Individual preference (an element of the described set)	Group preference (a potential description)		
	$a\mathbf{P}_G b$	$a\mathbf{I}_G b$	$b\mathbf{P}_G a$
$a\mathbf{P}_h b$	0	1	1
$a\mathbf{I}_h b$	1	0	1
$b\mathbf{P}_h a$	1	1	0

Given a potential group preferential ranking $\mathbf{R}_G$, one can treat the sum of the values of the functions of description errors for all the pairs of objects as a measure of how well $\mathbf{R}_G$ reflects the whole individual preferential ranking $\mathbf{R}_h$. This can be written for both functions, and for other, similar ones, as follows:

$$\ell(\mathbf{R}_h, \mathbf{R}_G) = \sum_{j=1}^{m-1} \sum_{t>j}^{m} \ell[e(a_j, a_i)]$$

For comparisons, one can use the mean value instead of the sum.

For a whole profile of individual preferential orderings $\mathbf{\Pi}_G$, we can use the mean value of the function of description errors for all individual preferential orderings in the profile as a measure of the descriptive adequacy of a potential group preferential ranking.

$$E[\ell(\mathbf{\Pi}_G, \mathbf{R}_G)] = \frac{1}{n} \sum_{h=1}^{n} \ell(\mathbf{R}_h, \mathbf{R}_G)$$

Let $\mathfrak{R}$ denote the set of potential group preferential rankings, i.e. order relations defined on the set of m objects. Its cardinality increases very rapidly with the number of ranked objects. This will be discussed in section 6.

An *optimal group preferential ranking*, i.e., one that best reflects the profile of individual preferential rankings, is a preferential ordering $\mathbf{R}^*_G$ such that its function of description errors has the lowest mean value

$$E[\ell(\mathbf{\Pi}_G, \mathbf{R}^*_G)] = \min_{\mathbf{R}_G \in \mathfrak{R}} E[\ell(\mathbf{\Pi}_G, \mathbf{R}_G)]$$

Naturally, an optimal group preferential ordering is not always uniquely specified, which is to say there may exist more than one preferential ordering that meets the condition above.

Below, I provide an example of how to determine the optimal group preferential ordering for a profile of preferential rankings of five people defined on a set of three objects, for both functions of description errors: ℓ_1 and ℓ_2.

Example 1

The profile of individual preferential rankings is as follows:

$$\begin{array}{ll} \mathrm{R}_1: & a\text{–}b \;\; c \\ \mathrm{R}_2: & b \;\; a \;\; c \\ \mathrm{R}_3: & b \;\; c \;\; a \\ \mathrm{R}_4: & b\text{–}c \;\; a \\ \mathrm{R}_5: & a\text{–}c \;\; b \end{array}$$

Tables 4 and 5 present the values of functions of description errors for all thirteen potential group preferential orderings as well as the functions' mean values. The reader may easily check the calculations herself. This allows one to identify the optimal group preferential ranking and the mean value of the function of description errors associated with it.

Table 4

The values of the function ℓ_1 of description errors associated with a description of individual preferential rankings and the determination of an optimal group ordering

Potential group preferential ranking $\mathbf{R}_G$	Individual preferential rankings					Mean value of the function of description errors $E[\ell_1(\mathbf{\Pi}_G, \mathbf{R}_G)]$
	$\mathbf{R}_1 =$ $a{-}b\ \ c$	$\mathbf{R}_2 =$ $b\ \ a\ \ c$	$\mathbf{R}_3 =$ $b\ \ c\ \ a$	$\mathbf{R}_4 =$ $b{-}c\ \ a$	$\mathbf{R}_5 =$ $a{-}c\ \ b$	
$a\ \ b\ \ c$	1	2	4	5	3	3.0
$a{-}b\ \ c$	0	1	3	4	4	2.4
$b\ \ a\ \ c$	1	0	2	3	5	2.2
$\boldsymbol{b\ \ a{-}c}$	2	1	1	2	4	**2.0**
$b\ \ c\ \ a$	3	2	0	1	5	2.2
$b{-}c\ \ a$	4	3	1	0	4	2.4
$c\ \ b\ \ a$	5	4	2	1	3	3.0
$c\ \ a{-}b$	4	5	3	2	2	3.2
$c\ \ a\ \ b$	5	6	4	3	1	3.8
$a{-}c\ \ b$	4	5	5	4	0	3.6
$a\ \ c\ \ b$	3	4	6	5	1	3.8
$a\ \ b{-}c$	2	3	5	4	2	3.2
$a{-}b{-}c$	2	3	3	2	2	2.4

There is only one optimal group preferential ranking: $\boldsymbol{b\ \ a - c}$. The mean value of the function of description errors is **2,0**.

Table 5

The values of the function ℓ_2 of description errors associated with a description of individual preferential rankings and the determination of an optimal group ordering

Potential group preferential ranking $\mathbf{R}_G$	Individual preferential rankings					Mean value of the function of description errors $E[\ell_1(\mathbf{\Pi}_G, \mathbf{R}_G)]$
	$\mathbf{R}_1 =$ $a{-}b\ \ c$	$\mathbf{R}_2 =$ $b\ \ a\ \ c$	$\mathbf{R}_3 =$ $b\ \ c\ \ a$	$\mathbf{R}_4 =$ $b{-}c\ \ a$	$\mathbf{R}_5 =$ $a{-}c\ \ b$	
$a\ \ b\ \ c$	1	1	2	3	2	1.8
$a{-}b\ \ c$	0	1	2	3	3	1.8
$\boldsymbol{b\ \ a\ \ c}$	1	0	1	2	3	**1.4**

Potential group preferential ranking $\mathbf{R}_G$	Individual preferential rankings					Mean value of the function of description errors $E[\ell_1(\mathbf{\Pi}_G, \mathbf{R}_G)]$
	$\mathbf{R}_1 =$ $a{-}b\ c$	$\mathbf{R}_2 =$ $b\ a\ c$	$\mathbf{R}_3 =$ $b\ c\ a$	$\mathbf{R}_4 =$ $b{-}c\ a$	$\mathbf{R}_5 =$ $a{-}c\ b$	
$b\ a{-}c$	2	1	1	2	2	1.6
$\boldsymbol{b\ c\ a}$	2	1	0	1	3	**1.4**
$b{-}c\ a$	3	2	1	0	3	1.8
$c\ b\ a$	3	2	1	1	2	1.8
$c\ a{-}b$	2	3	2	2	2	2.2
$c\ a\ b$	3	3	2	2	1	2.2
$a{-}c\ b$	3	3	3	3	0	2.4
$a\ c\ b$	2	2	3	3	1	2.2
$a\ b{-}c$	2	2	3	2	2	2.2
$a{-}b{-}c$	2	3	3	2	2	2.4

In this case, there are two optimal group preferential rankings: $\boldsymbol{b\ a\ c}$ and $\boldsymbol{b\ c\ a}$. For both, the mean value of the function of description errors is **1.4**. This is the minimum mean value of the function.

The group preferential rankings have been determined in the same way as the basic description parameters of the distribution of a single variable that have an interpretation in terms of the optimal description of the distribution of that variable. In classical methods of statistical description, error is defined as the difference of the true value and the value used to describe it, or $e(x, a) = x - a$.

The typical and most commonly used functions of description error estimation include:

1. The binary error function

$$\ell[e(x,a)] = \begin{cases} 0 & \text{if } x = a \\ 1 & \text{if } x \neq a \end{cases}$$

This is the simplest error function, which does not depend on the magnitude of the error, but only on whether or not an error has occurred.

2. Absolute value error function

$$\ell[e(x,a)] = |x - a|$$

This is a symmetric error function that estimates the consequences of error proportionally to its absolute value.

3. Square error function

$$\ell[e(x,a)] = (x-a)^2$$

This is a symmetric error function that is lenient when the errors are small, but punishing when the errors are large.

An optimal description is the number a^* such that it minimizes the mean value of the error function. The mean value of the function of description errors as estimated by a^*, i.e., $E[\ell(X,a^*)]$, is a measure of the quality of the optimal description.

$$E[\ell(X,a^*)] \leq E[\ell(X,a)] \text{ for any value of } a \in R$$

Central tendency parameters are optimal descriptions of a statistical variable with a fixed method of error estimation, whereas *dispersion parameters* are the mean values of error functions for an optimal description. Depending on which error function is used, central tendency parameters and dispersion parameters are as listed in table 6.

Table 6

Interpretation of central tendency and dispersion parameters in terms of an optimal description

Function of description errors $\ell[e(x,a)]$	Central tendency parameter	Dispersion parameter
	Optimal description a^*	Mean value of the function of optimal description errors $E\{\ell[e(X,a^*)]\}$
Binary	$Mo(X)$ Mode	$b(X)$ Probability of mode error
Absolute value	$Me(X)$ Median	$d(X)$ Mean absolute deviation
Square	$E(X)$ Mean	$D^2(X)$ Variance

4. Distances between preferential rankings

In the previous section, I discussed a method for statistical description of a set of individual preferences in terms of a group preferential ranking that best captures the underlying individual preferences. I also introduced a measure of the quality of such a description, i.e. of how well the group preferential ranking reflects the individual preferences. The problem of aggregating individual preferences into a group preference has long been the subject of analysis for social choice theorists.

At the end of the eighteenth century, in France, Marquis de Condorcet proposed that group decisions be made by comparing pairs of objects using the majority rule (five hundred years before Condorcet, the same method was contemplated by Ramon Liull – see McLean 1990). Although Condorcet's main purpose was to select the best object, rather than rank objects from best to worst, researchers who have investigated his writings (such as Guilbaud 1952 and Monjardet 2008) suggest that he also analyzed rankings (see also Balinski & Laraki 2010, chapter 4). More specifically, Condorcet believed that the best ranking is one with the largest sum of support votes in the preference relation for all pairs of objects. In such an ordering, if there exists an object that wins with every other object in pairwise comparisons (a so-called *Condorcet-winner*) then it comes first in the ranking; correspondingly, if there is an object that loses with every other object in pairwise comparisons (a so-called *Condorcet–loser*) then it comes last. However, it is hard to decide which of the two functions of description errors discussed in the previous section, ℓ_1 or ℓ_2, is closer to what Condorcet had in mind, for he only considered strict preferences, i.e. preferences without any indifferences.

It is believed today (e.g. by Young & Levenglick 1978, Young 1988, Meskanen & Nurmi 2006) that the method of aggregating individual preferences closest to Condorcet's intentions is based on a measure of distances between preferential orderings that was first proposed and later axiomatized by John G. Kemeny (1959, Kemeny & Snell 1962, chapter 2).

Kemeny represented the preference relation in terms of an antisymmetric matrix $\mathbf{K}$. The elements of the matrix are numbers $k_{jt}^h \in \{-1, 0, 1\}$ and

$$k_{jt}^h = \begin{cases} 1 & \text{if } a_j P_h a_t \\ 0 & \text{if } a_j I_h a_t \\ -1 & \text{if } a_t P_h a_j \end{cases}$$

By *the measure of the Kemeny distance* between two preferential rankings $\mathbf{R}_h$ and $\mathbf{R}_G$ of the same set of objects $\mathbf{A}$ we mean

$$d_K(\mathbf{R}_h, \mathbf{R}_G) = \frac{1}{2} \sum_{j=1}^{m} \sum_{t=1}^{m} \left| k_{jt}^h - k_{jt}^G \right|$$

where k_{jt}^G and k_{jt}^G denote the *jt*-th element of the matrix representing the two preferences $\mathbf{R}_h$ and $\mathbf{R}_G$.

The measure of the Kemeny distance between rankings is the only metric satisfying the following axioms:

(A.1.1) $d(\mathbf{R}_h, \mathbf{R}_g) \geq 0$, where the equality obtains iff rankings $\mathbf{R}_h$ and $\mathbf{R}_g$ are identical.

(A.1.2) $d(\mathbf{R}_h, \mathbf{R}_g) = d(\mathbf{R}_g, \mathbf{R}_h)$.

(A.1.3) $d(\mathbf{R}_h, \mathbf{R}_g) + d(\mathbf{R}_g, \mathbf{R}_i) \geq d(\mathbf{R}_h, \mathbf{R}_i)$, where the equality obtains iff ranking $\mathbf{R}_g$ lies between rankings $\mathbf{R}_h$ and $\mathbf{R}_i$.

(A.2) If ranking $\mathbf{R}'_h$ can be obtained by permutation from $\mathbf{R}_h$ and ranking $\mathbf{R}'_g$ can be obtained from $\mathbf{R}_g$ by the same permutation then $d(\mathbf{R}'_h, \mathbf{R}'_g) = d(\mathbf{R}_h, \mathbf{R}_g)$.

(A.3) If two rankings $\mathbf{R}_h$ and $\mathbf{R}_g$ are in agreement except for a certain subset $\mathbf{S}$ of k elements, which is a segment of both rankings, then it is possible to calculate $d(\mathbf{R}_h, \mathbf{R}_g)$ on the basis of rankings of elements belonging to that segment.

(A.4) The minimum positive distance is 1.

The first three axioms provide the basic constraints on distance. The last axiom merely establishes, in an arbitrary manner, a unit of measurement. Axioms (A.2) and (A.3) are decisive however. Other axiomatizations of this measure have also been developed (e.g., by Lissowski & Swistak 1995). It is sometimes called a city metric because its axiomatic definition contains a condition that presupposes the relation of "lying between" (A.1.3).

The Kemeny distance between two preferential rankings $\mathbf{R}_h$ and $\mathbf{R}_G$ is equal to the value of the error function ℓ_1 with $\mathbf{R}_h$ as its domain and $\mathbf{R}_G$ as its range

$$d_K(\mathbf{R}_h, \mathbf{R}_G) = \ell_1(\mathbf{R}_h, \mathbf{R}_G)$$

Kemeny distances between preferential rankings of three objects can be represented in the form of a distance map in figure 1. The smallest distance in this picture and the distance between two arbitrary orderings is defined as the length of the shortest path between them (along the line segments that connect them). Note that the distances are the same as in table 4. Note also that the distances involved cannot be represented faithfully on a plane, so the map is only a simplified illustration. A similar figure of distances between preferential rankings of four objects, reduced to strict linear preferences, is presented by Burak Can (2014: 115).

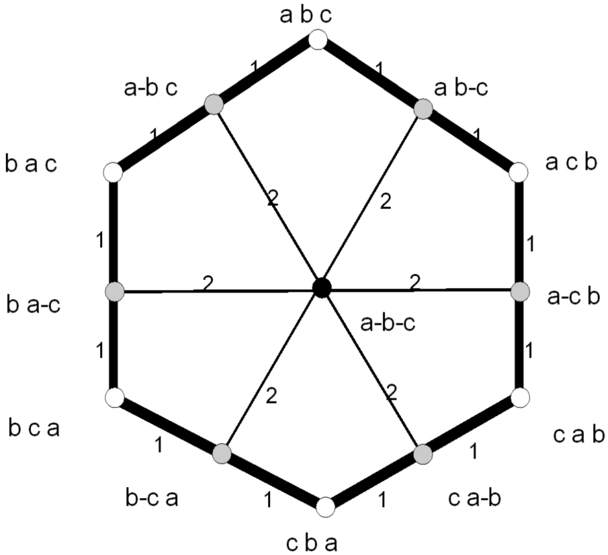

Figure 1. Kemeny distances d_K between all rankings of three objects

Kemeny proposed two procedures for determining a group preferential ordering. The first method consists in selecting ranking R_G^* such that the sum of distances from it is the smallest $\sum_{h=1}^{n} d_K(R_h, R_G^*)$. Such a group ordering is called the *median of individual preferences* and it is very often used in social choice theory. The second method consists in selecting ordering R_G^{**}, called the *mean of individual preferences*, that minimizes the sum of the squares of the distances $\sum_{h=1}^{n} d_K(R_h, R_G^{**})^2$. It is rarely used.

The minimum mean value of the function of description errors ℓ_1 associated with all individual preferential rankings in a given profile $\mathbf{\Pi}_G$ as captured by group preference R_G^* is equal to the mean Kemeny distance from the median of the individual preferences R_G^*.

$$\min_{\mathbf{R}_G \in \mathfrak{R}} E[\ell_1(\mathbf{\Pi}_G, \mathbf{R}_G)] = E[\ell_1(\mathbf{\Pi}_G, \mathbf{R}_G^*)] = \frac{1}{n}\sum_{h=1}^{n} \ell_1(\mathbf{R}_h, \mathbf{R}_G^*)$$
$$= \frac{1}{n}\sum_{h=1}^{n} d_K(\mathbf{R}_h, \mathbf{R}_G^*)$$

The choice of an optimal preferential ranking according to the function ℓ_1 of description errors is the same as that determined by Kemeny's method.

One can determine the value of the Kemeny distance between preferential rankings $\mathbf{R}_h$ and $\mathbf{R}_G$ directly on the basis of the order relations, represented as a matrix in section 2. The Kemeny distance is equal to the *symmetric difference* between the relations.

$$d_K(\mathbf{R}_h, \mathbf{R}_G) = |\mathbf{R}_h \cup \mathbf{R}_G| - |\mathbf{R}_h \cap \mathbf{R}_G| = |\mathbf{R}_h - \mathbf{R}_G| + |\mathbf{R}_G - \mathbf{R}_h|$$

For relations $\mathbf{R}_1$ and $\mathbf{R}_2$ specified in section 2, the value of the measure is:

$$d_K(\mathbf{R}_1, \mathbf{R}_2) = 15 - 7 = 4 + 4 = 8$$

Until recently, the measure of distance associated with the function of description errors ℓ_2 has not been discussed in the social choice literature (see Elkind, Faliszewski & Slinko 2012, Erdamar 2013, Nurmi 2014). Although published by Richard W. Hamming as early as 1950 (i.e., earlier than the Kemeny distance metric), its main applications were restricted to coding and signal transmission theories. The purpose of the measure was to estimate errors (compare codes or signals). Applied to comparing preference relations (preferential orderings), it shows differences between the preference relations under comparison. This measure of distance between preferential rankings satisfies all Kemeny axioms except (A.3).

We say that there is a difference between preferences $\mathbf{R}_h$ and $\mathbf{R}_G$ when the relation obtains between a given pair of objects in one of the preferences, but not in the other. A measure of the difference between two preferences, defined on a single set of objects, is the number of pairs of objects for which there is a difference between the two preferences. This measure is called *the measure of the Hamming distance.*

$$\begin{aligned} d_H(\mathbf{R}_h, \mathbf{R}_G) = |\{(a_i, a_j) \in A \times A : [(a_i, a_j) \in \mathbf{R}_h \wedge (a_i, a_j) \notin \mathbf{R}_G] \\ \vee [(a_i, a_j) \notin \mathbf{R}_h \wedge (a_i, a_j) \in \mathbf{R}_G]\}| \end{aligned}$$

Naturally,

$$d_H(\mathbf{R}_h, \mathbf{R}_G = \ell_2(\mathbf{R}_h, \mathbf{R}_G)$$

The minimum mean value of the function of description errors ℓ_2 associated with all individual preferential rankings in profile $\mathbf{\Pi}_G$ as captured by a group preferential ranking R^*_G is equal to the mean Hamming distance from the preferential ordering R^*_G.

$$\begin{aligned} \min_{\mathbf{R}_G \in \mathfrak{R}} E[\ell_2(\mathbf{\Pi}_G, \mathbf{R}_G)] = E[\ell_2(\mathbf{\Pi}_G, \mathbf{R}^*_G)] = \frac{1}{n}\sum_{h=1}^{n} \ell_2(\mathbf{R}_h, \mathbf{R}^*_G) \\ = \frac{1}{n}\sum_{h=1}^{n} d_H(\mathbf{R}_h, \mathbf{R}^*_G) \end{aligned}$$

Figure 2 shows a Hamming distance map d_H for preferential rankings of three objects. The distances are the same as in table 5. It is clear that they cannot be adequately represented on a plane.

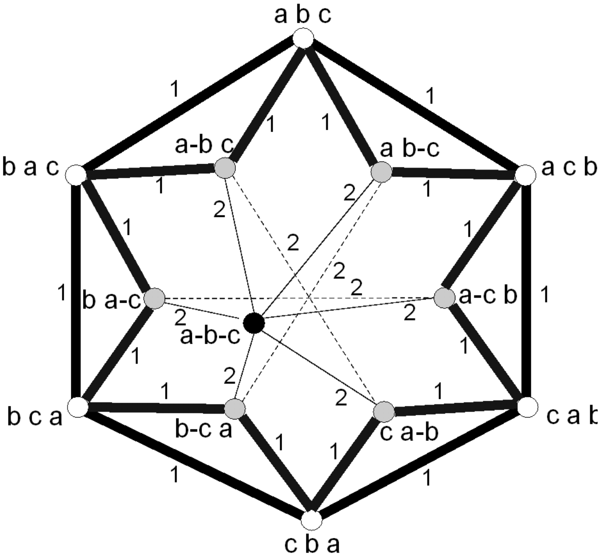

Figure 2. Hamming distances d_H between all rankings of three objects

One can determine the value of the Hamming distance between preferential orderings $\mathbf{R}_h$ and $\mathbf{R}_G$ directly on the basis of the ordering relations, displayed in the form of a matrix as in section 2.

$$d_H(\mathbf{R}_h, \mathbf{R}_G) = |\mathbf{R}_h| + |\mathbf{R}_G| - |\mathbf{R}_h \cup \mathbf{R}_G|$$

For relations $\mathbf{R}_1$ and $\mathbf{R}_2$ specified in section 2, the value of this measure of distance is:

$$d_H(\mathbf{R}_1, \mathbf{R}_2) = 11 + 11 - 15 = 7$$

5. A measure of the dependence of preferential rankings on a categorical variable

The measure of the dependence of preferential rankings on categorical variable Y assuming p values ($Y = \{1, 2, \ldots, k, \ldots, p\}$) is constructed in much the same manner as the measures involving statistical variables. Bear in mind that the construction of the measures of statistical dependence is based on the account of the pragmatic value of information. A general measure of the intensity of the statistical dependence of variable X on variable Y is defined as index $\Psi_{X|Y}$, which is a proportion of the "profit", or the reduction of the mean value of the function of description errors ℓ (Lissowski 1974b, 1976, 1977, Lissowski, Haman & Jasiński 2011, vol. 2).

The measure of the intensity of the statistical dependence of statistical variable X on variable Y when the function of description errors is ℓ

$$\Psi_{X|Y} = \frac{E[\ell(X, a^*)] - E\{\ell[(X, d^*(Y)]\}}{E[\ell(X, a^*)]}$$

where a^* is the optimal description of variable X when the value of Y is ignored, and d^* is the optimal description of variable X with information concerning the value of variable Y taken into account.

The measure of the dependence of preferential rankings on categorical variable Y when the function of description errors is ℓ

$$\mu_{\Pi_G|Y} = \frac{E[\ell(\Pi_G, R^*_G)] - E[\ell(\Pi_{G|Y}, R^*_{G|Y}|Y)]}{E[\ell(\Pi_G, R^*_G)]}$$

where $R^*_{G|Y}$ is a function that assigns, to every value of categorical variable Y, an optimal group preferential ordering for the profile of preferential rankings in the subset specified on the basis of that value of Y. Accordingly, $R^*_{G|Y}$ is a *generalized regression* with preferential orderings as its values.

A measure of the dependence of preferential orderings on variable Y tells us to what extent the mean value of the function of description errors is reduced when we know to which subset, specified on the basis of variable Y, the preferential ordering belongs. Like all measures of the intensity of statistical dependence, this measure assumes the values from the interval $[0, 1]$. Its value is 0 when the group preferential rankings in the subsets are identical, and 1 when the description of preferential orderings in those subsets, based on generalized regression and on information about the value of Y, contains no error.

Example 2

Suppose for the sake of illustration that the set consists of twenty people who rank three objects: a, b and c. The set is divided into three subsets on the basis of the value of variable Y. The first subset ($Y = 1$) is a group of five people whose profile of preferences has been analyzed in example 1. The joint distribution of preferential rankings and variable Y is displayed in table 7.

Table 7

The joint distribution of preferential rankings and categorical variable *Y*

Preferential ranking R_h	Value of variable Y			Total
	1	2	3	
$a\ b\ c$		1		1
a–$b\ c$	1	1		2
$b\ a\ c$	1	2		3
$b\ a$–c				0
$b\ c\ a$	1	2		3
b–$c\ a$	1	3	1	5
$c\ b\ a$		1		1
$c\ a$–b				0
$c\ a\ b$			1	1
a–$c\ b$	1		2	3
$a\ c\ b$			1	1
$a\ b$–c				0
a–b–c				0
Total	5	10	5	20

If we assume the function of description errors ℓ_1 and, using the method described above, determine the optimal group preferential rankings along with the corresponding mean values of the functions of description errors for the subsets specified on the basis of variable Y, we obtain the following values.

Table 8

Optimal group preferential rankings and the corresponding mean values of the function of description errors ℓ_1 in subsets specified on the basis of variable *Y*

	Subsets specified on the basis of the value of variable Y			Whole set
	1	2	3	
Optimal group preferential ranking	$b\ a$–c	$b\ c\ a$	a–$c\ b$	b–$c\ a$
Mean value of the function of of description errors ℓ_1	2.0	1.6	1.6	2.3
Number of rankings	5	10	5	20

Table 8 depicts a generalized regression of preferential orderings with respect to categorical variable Y as well as the mean values of the function of description errors. The mean value of the description errors associated with this regression is equal to 1.7.

$$E[\ell_1(\Pi_{G|Y}, R^*_{G|Y}|Y)] = \frac{2.0 \times 5 + 1.6 \times 10 + 1.6 \times 5}{20} = \frac{34}{20} = 1.7$$

The measure of the dependence of preferential orderings on categorical variable Y when the function of description errors is ℓ_1 is equal to 0.261.

$$\mu^1_{\Pi_G|Y} = \frac{E[\ell_1(\Pi_G, R^*_G)] - E[\ell_1(\Pi_{G|Y}, R^*_{G|Y}|Y)]}{E[\ell_1(\Pi_G, R^*_G)]} = \frac{2.3 - 1.7}{2.3} = \frac{0.6}{2.3} = 0.261$$

Information about the value of categorical variable Y has reduced the mean value of the function of description errors by 0.6, or 26.1% of the value the function of description errors had when the information was not available.

Given ℓ_2 as the function of description errors, the determination of the generalized regression of the preferential orderings with respect to categorical variable Y and the mean values of the function of description errors ℓ_2 proceeds along very similar lines. In this case, the generalized regression of the preferential orderings is not uniquely specified. There are four equally optimal generalized regressions. The mean value of the corresponding function of description errors is 1.2.

Table 9

Optimal group preferential rankings and the corresponding mean values of the function of description errors ℓ_2 in subsets specified on the basis of the value of variable Y

	Subsets specified on the basis of the value of variable Y			Whole set
	1	2	3	
Optimal group preferential ranking	$b\ a\ c$ $b\ c\ a$	$b\ c\ a$	$c\ a\ b$ $a{-}c\ b$	$b\ c\ a$
Mean value of the function of of description errors ℓ_2	1.4	1.1	1.2	1.45
Number of rankings	5	10	5	20

$$E[\ell_2(\Pi_{G|Y}, R^*_{G|Y}|Y)] = \frac{1.4 \times 5 + 1.1 \times 10 + 1.2 \times 5}{20} = \frac{24}{20} = 1.2$$

The measure of the dependence of preferential rankings on categorical variable Y when the function of description errors is ℓ_2 is equal to 0.172.

$$\mu^2_{\Pi_G|Y} = \frac{E[\ell_2(\Pi_G, R^*_G)] - E[\ell_2(\Pi_{G|Y}, R^*_{G|Y}|Y)]}{E[\ell_2(\Pi_G, R^*_G)]} = \frac{1.45 - 1.2}{1.45} = \frac{0.25}{1.45} = 0.172$$

Information about the value of categorical variable Y has reduced the mean value of the function of description errors by 0.25, or 17.2% of the mean value the function of description errors assumed when the information was unavailable. In this case, the mean value of the error function is equal to the frequency of (any type of) error.

In a similar way, one can determine the regressions of preferential rankings with respect to a categorical variable as well as a measure of the intensity of this dependence for any function of description errors one chooses.

6. Computational complexity

The number of all possible preferential rankings $\mathfrak{R}$, as well as of preferential orderings without indifference, increases rapidly as a function of the number of objects m. For example, table 10 displays the numbers of such rankings for $m = 2, \ldots, 6$.

The determination of the optimal group preferential ranking, the corresponding mean value of the function of description errors, the regression of preferential rankings and the measure of the intensity of the dependence on a categorical variable is a task of significant computational complexity. Of course, when the number of objects is small, the required calculations are fairly easy to perform. This is not so when the number of objects is large.

Table 10

The number of all preferential rankings and rankings without indifference for m objects

Number of objects	m	2	3	4	5	6
Number of all preferential rankings	M	3	13	75	541	4683
Number of preferential rankings without indifference	$m!$	2	6	24	120	720

There is a direct solution to the problem in the case when, for all pairs of objects, the group preference relations that optimally reflect the individual preferences in the profile of individual preferences (i.e., those with the smallest sums of the functions of description errors) satisfy the condition of transitivity, and so determine an order relation on the set of objects. Unfortunately, this rarely happens and, often, the search for an optimal solution is very complex indeed.

A number of algorithms have been developed for finding a group preferential ordering that minimizes the sum of Kemeny distances. There are many publications addressing this problem, for a review of the literature see, for example, Charon & Hudry (2010).

Translated by Witold Hensel

NOTE

[1] An earlier version of this paper was presented at the conference *New Approaches in Quantitative Analysis in the Social Sciences*, September 26–28, 2012, Jabłonna, Poland.

REFERENCES

Balinski, Michel, Laraki, Rida. (2010). *Majority Judgment: Measuring, Ranking, and Electing.* Cambridge: The MIT Press.

Can, Burak. (2014). Weighted distances between preferences. *Journal of Mathematical Economics 51*, 109–115.

Charon, Irene, Hudry, Olivier. (2010). An updated survey on the linear ordering problem for weighted or unweighted tournaments. *Annals of Operation Research 175*, 107–158.

Elkind, Edith, Faliszewski, Piotr, Slinko, Arkadii. (2012). Rationalizations of Condorcet-consistent rules via distances of hamming type. *Social Choice and Welfare 39*, 891-905.

Erdamar, Bora. (2013). *Informational Frameworks for Collective Decision Making: "A Suggested Compromise".* Economics and Finances. Ecole Polytechnique X. PhD Thesis.

Guilbaud, Georges-Théodule. (1952). Les théories de l'intérêt général et le problème logique de l'agrégation. *Economie Appliquée 5(4)*, 501–584. Complete English translation in *Electronic Journ@l for History of Probability and Statistics 4.1*, 2008.

Hamming, Richard W. (1950). Error Detecting and Error Correcting Codes. *The Bell System Technical Journal 29*, 147–160.

Kemeny, John G. (1959). Mathematics without numbers. *Daedalus 88*, 577–591.

Kemeny, John G., Snell, Laurie. (1962). *Mathematical Models in the Social Sciences.* New York: Ginn.

Lissowski, Grzegorz. (1974a). Statystyczny opis zbioru uporządkowań preferencyjnych (Statistical description of a set of preferential rankings). *Prakseologia Nr 3–4 (51-52)*, 379–413.

Lissowski, Grzegorz. (1974b). Statistical laws and prediction. *The Polish Sociological Bulletin No 2*, 23–37.

Lissowski, Grzegorz. (1976). Interpretation of statistical measures. *The Polish Sociological Bulletin. Special issue. Studies in Methodology*, 89–111.

Lissowski, Grzegorz. (1977). Statistical association and prediction. In K. Szaniawski (ed.), *Problems of Formalization in the Social Sciences.* 217–245. Warszawa: Ossolineum.

Lissowski, Grzegorz. (2000). Metody agregacji indywidualnych preferencji (The methods of aggregation of individual preferences). *Studia Socjologiczne No 1–2 (156–157)*, 79-103.

Lissowski, Grzegorz, Haman, Jacek, Jasiński, Mikołaj. (2011). *Podstawy statystyki dla socjologów. Tom 1. Opis statystyczny, Tom 2. Zależności statystyczne, Tom 3. Wnioskowanie statystyczne.* (Fundamentals of Statistics for Sociologists. Vol. 1. Statistical Description, Vol. 2. Statistical Association, Vol. 3. Statistical Inference). Seria Wykłady z Socjologii. Warszawa: Wydawnictwo Naukowe "Scholar". Second edition.

Lissowski, Grzegorz, Swistak, Piotr. (1995). Choosing the best social order: new principles of justice and normative dimensions of choice. *American Political Science Review 89*, 74–96.

Meskanen, Tommi, Nurmi Hannu. (2006). Distance from consensus: a theme and variations. In B. Simeone and P. Pukelsheim (eds.), *Mathematics and Democracy. Recent Advances in Voting Systems and Collective Choice.* 117–132. Berlin: Springer.

McLean, Iain. (1990). The Borda and Condorcet principles: Three Medieval applications. *Social Choice and Welfare 7*, 99–108.

Monjardet, Bernard. (2008). "Mathématique Sociale" and mathematics. A case study: Condorcet's effect and medians. *Electronic Journ@l for History of Probability and Statistics 4.1.*

Nurmi, Hannu. (2014). Are we done with preference rankings? If we are, then what? http://www.utu.fi/yksikot/soc/yksikot/pcrc/Documents/Nurmi_oct2014.pdf Internet: 26-02-2015.

Young, H. Peyton. (1988). Condorcet's theory of voting. *American Political Science Review 82*, 1231–1244.

Young, H. Peyton, Levenglick, A. (1978). A consistent extension of Condorcet's election principle. *SIAM Journal on Applied Mathematics 35 (2)*, 285–300.