Classifying Behaviors in Risky Choices

Kontek, Krzysztof

Artal Investments

12 July 2010

Classifying Behaviors in Risky Choices

Krzysztof Kontek[1]
Artal Investments, Warsaw[2]

Abstract

This paper presents a nonparametric approach to classification of data from lottery experiments. Using very basic mathematical tools the paper endeavors to answer the questions: How to determine the "average" subject in a group? How to find a subject presenting the most similar behavior to a given one? How to detect outlier subject(s)? How to classify behaviors by their dissimilarity from the perfectly rational decision making? How to rank subjects by risk attitudes? How to cluster subjects? This paper demonstrates that the answer to all of these questions may be found non-parametrically, without the use of any specific model.

JEL classification: C02, C14, C81, C91, D03, D81

Keywords: Lottery experiments, Certainty Equivalents, Risk Attitude, Cluster Analysis, Non-parametric Methods, Relative Utility Function.

1. Introduction

1.1. There is an enormous number of theories explaining lottery experiments (and more generally risky choices), which were presented during the last 50 years. The most prominent of them is Prospect Theory for which Daniel Kahneman was awarded the Nobel Prize for economics in 2002. Despite this, it is hard to find any theory which is able to give simple answers to such questions as which subject presents the "average" behavior within a group or which subject is the most risk averse or risk seeking? One of the reasons for that is that the models typically use several functions and parameters (like the value and probability weighting functions in the case of Prospect Theory) and the resulting behavior or risk attitude pattern is described by a combination of them.

This, however, seems to be a very unfortunate situation as the questions posed appear to be very basic. Besides the sample "average" or extreme values may be found in most of

[1] The author is grateful to Stefan Traub from the Department of Economics, University of Bremen and Ulrich Schmidt from the Department of Economics, University of Kiel, for making the results of their experiments available.
[2] Contact: ul. Chrościckiego 93/105, 02-414 Warsaw, Poland,
e-mail: kontek@artal.com.pl, kkontek2000@yahoo.com.

economic problems non-parametrically, i.e. even without involving any specific model.

This paper is intended to find answers to several important questions regarding a subject's behavior in risky choices. As the mathematical tools needed for this purpose are very basic, so it is the "language" of this paper. Point 2 presents a very basic introduction into distance measures, which is the basic tool used throughout the paper. Point 3 presents experimental data, which serve to demonstrate examples in the further part of this research. Point 4 is devoted to determining the "average" subject within a group. Point 5 demonstrates how to find a subject presenting the most similar behavior to a given one. Point 6 is devoted to detecting outlier subjects. Point 7 demonstrates the method of ordering subjects by dissimilarity from the perfectly rational decision maker. Ranking subjects by risk attitude is presented in Point 8. Point 9 is devoted to clustering the subjects. Point 10 summarizes the study with the conclusion that all of the presented questions and problems may be answered non-parametrically, without involving any specific model.

In order to check the obtained results with a parametrical approach the paper presents in the Appendix individual relative utilities (Kontek, 2010) derived from the examined data-set. Comparison of the results obtained non-parametrically and using the relative utility model shows their strong correspondence and validates both methods.

2.1. Basic Introduction into Distance Measures

2.1. An "average" value of a given data set may be defined in several ways. The most common are mean and median. The median is typically preferred as it is more robust to outliers than the mean. No complex theory is required to use mean and median in practice, but it is worth restating that their use as measures of central tendency is justified by distance measures.

2.2. Assuming there is a vector of data $x = \{x_1, x_2, .., x_n\}$ one is looking for such x_{mean} value, which minimizes the (Squared) Euclidean Distance:

$$S_{mean} = \sum_{i=1}^{n} \left(x_i - x_{mean} \right)^2. \tag{2.1}$$

The minimum distance may be found by comparing the derivative of (2.1) to zero:

$$\frac{dS_{mean}}{dx_{mean}} = 2\sum_{i=1}^{n} \left(x_i - x_{mean} \right) = 0, \tag{2.2}$$

which results in

$$x_{mean} = \frac{\sum_{i=1}^{n} x_i}{n}, \tag{2.3}$$

which is the mean value of all sample data.

2.3. That the use of median results from minimizing the Manhattan Distance is less obvious. The Manhattan distance is expressed as a sum of absolute differences rather than of their squares:

$$S_{med} = \sum_{i=1}^{n} |x_i - x_{med}|. \tag{2.4}$$

In the ordered sample of x_i, (2.4) may be presented as:

$$S_{med} = \sum_{i=1}^{k} (x_{med} - x_i) + \sum_{i=k+1}^{k+m} (x_i - x_{med}). \tag{2.5}$$

The derivative of (2.5) is

$$\frac{dS_{med}}{dx_{med}} = k - m, \tag{2.6}$$

which assumes the value of 0 when the number of data points to the left equals the number of points to the right. As the second derivative of (2.5) is always 0, (2.5) has no curvature and is a straight line between consecutive points. It follows that (2.5) assumes the minimum at the middle point for an odd sample and at any point between two middle points for an even sample (usually the mean of middle points is assumed). This is exactly the median definition.

2.4. The basic procedures presented in Points 2.2. and 2.3. need to be slightly modified if the solution is sought among data points. It is because the value of x_{mean} (2.3) may be any point in the domain, not necessarily one of the data points. In order to determine such a solution (called "medoid" instead of "mean" or "centroid") it is necessary to calculate the Euclidean Distance (2.1) for each data point and to select this one for which the distance assumes the minimum value.

There is also a slight difference in the case of determining the median of an even sample. As the solution is sought among data points and the minimized function (2.4) assumes a constant value between the middle points it must be accepted that there are two median values in this case.

2.5. The data points may, in general, be vectors of data. The procedure of determining the "average" values in this case will be shown after presenting the data from lottery experiments.

3. Data Set

3.1. The data set used in this research is the experimental data presented by Traub and Schmidt (2009), whose research concerned the relationship between WTP (Willingness to Pay) and WTA (Willingness to Accept)[3]. Twenty-four subjects participated in the experiment. A subset of this data concerning certainty equivalents is presented in Table 3.1.

Lottery	Outcomes	Probabilities	1	2	3	.	22	23	24
.	.	.	.	.	.	.	.	.	.
2	0, 30	0.75, 0.25	10	7	7.5	.	7.5	7.5	7
3	0, 10, 30	0.3, 0.6, 0.1	8.7	10	9	.	9	9	20
4	10, 30, 40	0.6, 0.1, 0.3	16	20	21	.	21	25	20
.	.	.	.	.	.	.	.	.	.
56	0, 10, 40	0.25, 0.35, 0.4	17	25	19.5	.	19.5	20	19

Table 3.1. The subset of experimental data. Data in columns 1..24 present certainty equivalents defined by respective subjects.

3.2. Each column 1..24 represents a vector of certainty equivalents defined by respective subjects. According to the methodology presented in Point 2 the task of determining the "average" subject is to find the mean and median vectors of certainty equivalents. However, as the examined lotteries have different minimum and maximum values, the certainty equivalents assume values in different ranges. This complicates calculation of distance minimum, as lotteries with a broader outcome range would have a greater impact on the result than lotteries with narrower range of outcomes. Therefore, it is quite natural to normalize the lottery outcomes and respective certainty equivalents using:

$$r = \frac{ce - P_{\min}}{P_{\max} - P_{\min}}, \tag{3.1}$$

where r denotes the normalized certainty equivalent, ce denotes the certainty equivalent, $P_{\min}$ denotes the minimum lottery outcome, and $P_{\max}$ denotes the maximum lottery outcome.

The data presented formerly in Table 3.1. are now shown in Table 3.2. Both lottery outcomes and respective certainty equivalents have been transposed to the [0,1] interval thus equalizing the impact of each lottery on the minimization procedure.

Lottery	Outcomes	Probabilities	1	2	3	.	22	23	24
	.	.	.	.	.	.	.	.	.
2	0, 1	0.75, 0.25	1/3	7/30	¼	.	¼	1/4	7/30
3	0, 1/3, 1	0.3, 0.6, 0.1	29/100	1/3	3/10	.	3/10	3/10	2/3
4	0, 2/3, 1	0.6, 0.1, 0.3	1/5	1/3	11/30	.	11/30	1/2	1/3
	.	.	.	.	.	.	.	.	.
56	0, ¼, 1	0.25,0.35, 0.4	17/40	5/8	39/80		39/80	1/2	19/40

Table 3.2. The normalized subset of experimental data. Data in columns 1..24 represent normalized certainty equivalents given by respective subjects.

[3] The data were also analyzed in another paper by Hey, Morone and Schmidt (2009).

Instead of presenting the whole data set in a table form, it is demonstrated as individual relative utilities (Kontek, 2010). This is done in Appendix, as the purpose of this paper is to consider solely a nonparametric approach. However the reader is encouraged to compare these utilities with the results obtained in the remaining part of this paper.

4. Determining the "Average" Subject

4.1. Determining the "average" data point using distance measures (2.1) and (2.4) has to be modified as each data point is now a vector of normalized certainty equivalents. There are 24 vectors (as the number of subjects participating in the experiment) and their length is 54 (as the number of lotteries considered[4]). These vectors describe fully the behavior of subjects in the lottery experiment therefore they will be called "subjects" or "behaviors" in the remaining part of this paper.

The mean Manhattan Distance of a given subject to all other subjects is calculated using:

$$S_i = \frac{1}{(m-1)n} \sum_{j=1}^{m} \sum_{k=1}^{n} \left| r_{i,k} - r_{j,k} \right|, \tag{4.1}$$

where n denotes the number of lotteries, m denotes the number of subjects, i denotes the respective subject, and $r_{i,k}$ denotes the normalized certainty equivalent of ith subject in the kth lottery. The value m - 1 appears in the denominator as there are m - 1 distances considered for a given subject.

The mean Euclidean Distance of a given subject to all other subjects is calculated using:

$$S_i = \sqrt{\frac{1}{(m-1)n} \sum_{j=1}^{m} \sum_{k=1}^{n} \left(r_{i,k} - r_{j,k} \right)^2}. \tag{4.2}$$

The subjects are then ordered by respective distances. This ordering is presented in Table 4.1.

Manhattan Distance																						
20	21	22	3	7	1	4	12	18	2	24	11	23	5	10	14	17	15	9	19	13	16	6	8
Euclidean Distance																						
22	3	20	21	7	1	18	4	12	2	11	24	5	14	23	10	17	15	19	13	9	6	16	8

Table 4.1. Subjects ordered by the mean Manhattan and Euclidean Distances to other subjects. Smallest distances on the left.

The mean distances of respective subjects to all other subjects are presented graphically in Figure 4.1

[4] There were altogether 56 lotteries, but two were removed from this research, as they were lotteries with a certain payment (e.g. 30GBP with a probability of 1).

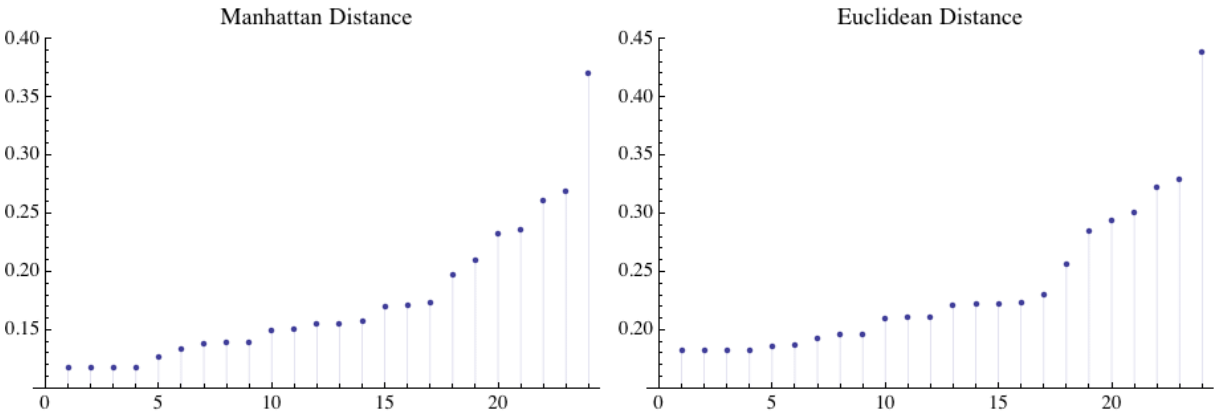


Figure 4.1. The mean Manhattan and Euclidean Distances of respective subjects to all other subjects. Ordering as in Table. 4.1.

The lowest mean Manhattan distance to other subjects is that of subjects 20 and 21. Therefore they are the median subjects in the examined group.

The lowest mean Euclidean Distance to other subjects is that of subject 22. Therefore it is the mean subject in the examined group.

It should be noted that subjects 3 and 22 have similarly low mean distances to other subjects. The next in these rankings are subjects 1 and 7, meaning that their behaviors are also close to the average in the group.

5. Finding the Most Similar Subject(s)

5.1. One of the questions, which may be posed when analyzing lottery data is which subject presents the most similar behavior comparing to a given one. This problem is easily solved by calculating distances between pairs of subjects. Using Manhattan Distance results in:

$$S_{i,j} = \frac{1}{n}\sum_{k=1}^{n}\left|r_{i,k} - r_{j,k}\right|, \tag{5.1}$$

and using Euclidean Distance in:

$$S_{i,j} = \sqrt{\frac{1}{n}\sum_{k=1}^{n}\left(r_{i,k} - r_{j,k}\right)^2}, \tag{5.2}$$

where $S_{i,j}$ denotes the distance between *i*th and *j*th subject, *n* denotes the number of lotteries considered, and $r_{i,k}$ denotes the normalized certainty equivalent of *i*th subject in the *k*th lottery. Quite obviously the distances (5.1) and (5.2) determine the level of dissimilarity between subjects; lower the distance – more similar are behaviors presented by subjects. The most similar subject to a given subject *i* can therefore be found by determining the subject *j,* which

is the nearest to subject *i*. The results for each subject using the Manhattan Distance are presented in Table 5.1

1	2	3	4	5	6	7	8	9	10	11	12
3	3	22	20	22	14	20	14	1	19	3	3
0.071	0.080	0.003	0.075	0.104	0.155	0.033	0.323	0.166	0.101	0.092	0.056
13	14	15	16	17	18	19	20	21	22	23	24
19	12	1	11	10	1	10	21	20	20	20	3
0.102	0.093	0.149	0.170	0.122	0.071	0.101	0.000	0.000	0.001	0.078	0.087

Table 5.1. The nearest (most similar) subject by Manhattan Distance. Respective subjects are presented on white background, the nearest subjects on light blue, the mean distance between the neighbors on gray.

It should be noted that subjects 20 and 21 present the same behavior, as the distance between them is 0. Interestingly to note that they were determined in Point 4 as two median subjects in the group. Subject 20 appears to be the most similar to 5 other subjects. This is also the case with subject 3. This confirms the results obtained in Point 4 and proves that subjects 20, 21 and 3 present the behavior, which is common in the examined group.

The nearest subjects using Euclidean Distance are presented in Table 5.2.

1	2	3	4	5	6	7	8	9	10	11	12
3	7	22	7	7	14	20	14	1	4	14	3
0.092	0.104	0.012	0.097	0.157	0.204	0.053	0.381	0.251	0.133	0.128	0.065
13	14	15	16	17	18	19	20	21	22	23	24
10	12	12	15	7	1	10	21	20	20	7	4
0.138	0.126	0.200	0.229	0.149	0.097	0.147	0.000	0.000	0.007	0.119	0.129

Table 5.2. The nearest (most similar) subject by Euclidean Distance. Respective subjects are presented on white background, the nearest subjects on light blue, the average distance on gray.

It should first be noted that changing the measure (from Manhattan to Euclidean Distance) changes the nearest subject in 11 out of 24 cases. This leads to slightly different conclusions than in the previous case. Subject 22, which was determined as the mean subject, is the nearest only to subject 3. The most "popular" is subject 7, whose behavior is the most similar to 5 other subjects. Subject 20 is the most similar to 3 subjects[5].

These conclusions confirm to a big extent the results presented in Point 4. The differences may result from the fact that having similar mean distance to other subjects (Table 4.1.) does not necessarily mean that subjects are similar to each other (Table 5.1 and 5.2). This should be quite obvious when one imagines two points located on different parts of a circle, which have the same distance to the circle centre. The case considered here is definitely much more difficult to imagine as the subjects are described not in 2 but in 54 dimensions.

[5] The reader is encouraged here once again to compare the results presented in this Point with individual relative utilities presented in the Appendix.

6. Detecting Outlier Subject(s)

6.1. The results presented in Points 4 and 5 enable detection of outliers, i.e. subject demonstrating behaviors, which vary substantially from the average in the group. One may see from Table 4.1 that subject 8 is the most distant from all other members of the group, so it may be suspected of being an outlier. This observation is confirmed by the data in Tables 5.1 and 5.2. The distance of subject 8 to its most similar neighbor is large comparing with other pairs. This means that subject 8 is an obvious outlier and his/her data should be treated carefully if not even declassified from further considerations.

It should be noted that other subjects being quite distant from their nearest neighbors. These are subjects 9, 16, 6 and 15. Not surprisingly, the same subjects appear in Table 4.1 as the most distant from all subjects in the group. This shows that their behavior differs substantially from the average.

6.2. Detecting outliers may lead to declassifying data from further analysis. This, however, should be done very carefully and only in exceptional cases[6].

7. Ordering by Expected Value Dissimilarity

7.1. As shown in Point 6 it is possible to measure dissimilarity between behaviors presented by a pair of subjects. However a question may be posed how a specific behavior differs from that of the perfectly rational decision maker? The only difference in the approach would be that the dissimilarity is measured between the specific behavior and the expected value of the lottery, as this is the value, which is chosen by a perfectly rational decision maker. The calculation can be made using Manhattan Distance:

$$S_i = \frac{1}{n}\sum_{k=1}^{n}\left|r_{i,k} - EL_k\right|, \tag{7.1}$$

where EL_k denotes the normalized expected value of kth lottery. Similarly using the Euclidean Distance results in:

$$S_i = \sqrt{\frac{1}{n}\sum_{k=1}^{n}\left(r_{i,k} - EL_k\right)^2} \tag{7.2}$$

The obtained distances may be ordered giving the classification by the behavior dis-

[6] Hey et. al. (2009), who used the same data as in the present research, stated "*We omit two of the 24 subjects (subjects 20 and 21) who answered all questions as if they were perfect expected - value maximizers*". It is not quite clear whether they omit the subjects only in one of the tables or also in the further analysis. If so, it would mean that authors declassified subjects 20 and 21, which, as presented, are the median subjects in the examined group, but left subject 8, which is the obvious outlier.

similarity from the expected value maximization. This ordering is presented in Table 7.1

Manhattan Distance																						
20	21	22	3	7	12	1	4	23	2	18	24	11	5	14	17	10	15	9	13	19	16	6	8
Euclidean Distance																						
20	21	22	3	7	12	1	4	2	18	23	11	24	14	17	10	5	15	19	13	9	6	16	8

Table 7.1. Ordering of subjects by the behavior dissimilarity from the Expected Value Maximization. More "rational" subjects to the left.

The values of dissimilarity from the Expected Value maximization are presented graphically in Figure 7.1.

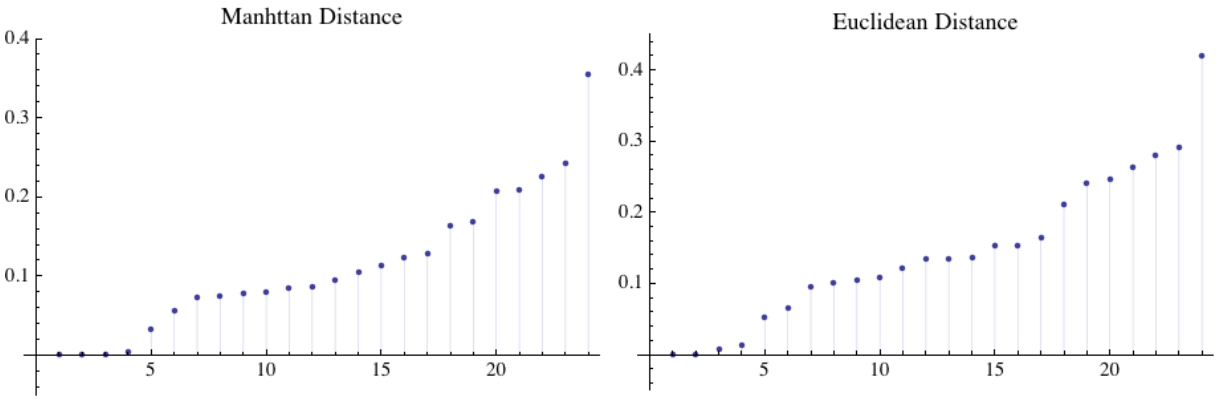


Figure 7.1. Behavior dissimilarity from the expected value maximization presented by respective subjects. Ordering of subjects as in Table 7.1.

As seen Subjects 20, 21, 22, and 3 present behavior, which is identical or almost identical with the expected value maximization. It means these subjects are perfectly rational decision-makers. On the other hand subject 8, which was detected earlier as the outlier, presents the behavior, which is the most dissimilar from the expected value maximization. This would mean that his/her behavior is far from rational.

8. Ordering by Risk Attitude

8.1. Let us now consider another method of classifying subjects according to their risk attitude. It should be quite apparent that the certainty equivalent expected for a given lottery becomes lower with the increasing risk aversion of the examined subject. It gives an indication that risk attitudes of two subjects may be compared using the following function:

$$S_{i,j} = \frac{1}{n}\sum_{k=1}^{n}\left(r_{i,k} - r_{j,k}\right) \tag{7.3}$$

where i and j denote respective subjects, k denotes a specific lottery, and n denotes the number of lotteries. Equation (7.3) has a similar form to (5.1) but, by no means, is any measure of distance as it may assume negative values as well. Instead, it defines the mean shift of nor-

malized certainty equivalents between two subjects. It should be seen that a negative value of *S* indicates that subject *i* is generally more risk averse than subject *j*, whereas a positive value would indicate that subject *i* is generally more risk seeking than subject *j*. The term "generally" is used here because (7.3) is calculated for a set of lotteries covering (hopefully) a wide range of risky choices.

8.2. Please note that *S* (7.3) is a relative measure as it compares certainty equivalents of two subjects. However (7.3) can be presented alternatively as:

$$S_{i,j} = \frac{1}{n}\sum_{k=1}^{n} r_{i,k} - \frac{1}{n}\sum_{k=1}^{n} r_{j,k} = RS_i - RS_j \tag{7.4}$$

i.e. as the difference between absolute measures of risk attitude presented by respective subjects. In fact RS_i and RS_j are the means of normalized certainty equivalents expected by a subject for a set of lotteries:

$$RS = \frac{1}{n}\sum_{k=1}^{n} r_k \tag{7.5}$$

The higher the risk aversion presented by a subject – the lower the value of *RS*. The higher the risk seeking presented by a subject – the higher the value of *RS*. Therefore this measure corresponds with the risk-seeking attitude of a given subject. The *RS* assumes the value of 0 when the subject is maximally risk averse (i.e. all certainty equivalents are equal to minimum lottery outcomes). *RS* assumes the value of 1 when the subject is maximally risk-seeking (i.e. all certainty equivalents are equal to maximum lottery outcomes). *RS* should assume the value of 0.5 when a subject is generally risk neutral (for instance when the subject is a perfect expected value maximizer). Due to the range of values assumed by RS it is straightforward to define the Risk Aversion measure as:

$$RA = 1 - RS \tag{7.6}$$

8.3. The values of *RS* and *RA* for a risk neutral person would assume the value of 0.5 only when the set of examined lotteries consists of lotteries with randomly distributed probabilities of their winning. Otherwise the result may be biased towards 0 or 1. Therefore it is better to use the Risk Seeking and Risk Aversion measures adjusted by lottery expected values:

$$RSA = \frac{1}{n}\sum_{k=1}^{n}\left(r_k - EL_k\right) \tag{7.7}$$

where EL_k denotes the normalized expected value of *k*th lottery. Expected values serve here only as constants, so they do not change preferences defined by *RS* (7.5). However they help

to avoid the measure bias and shift the result to a more convenient range of values. *RSA* assumes values in the range [-0.5, 0.5], and is negative for a generally risk averse person and positive for a generally risk seeking person. Adjusted Risk Aversion measure may be then symmetrically defined as:

$$RAA = -RSA = \frac{1}{n}\sum_{k=1}^{n}\left(EL_k - r_k\right) \tag{7.8}$$

It appears that the Adjusted Risk Aversion measure is the mean difference between normalized expected values and certainty equivalents. Synonymously, the person is risk averse when the certainty equivalents are on average lower than lottery expected values, and the person is risk seeking when the certainty equivalents are on average greater than lottery expected values[7].

8.4. It has to be added that the definition of the risk attitude presented above may lead to some interesting conclusions. For example that a risk neutral person does not necessarily have to be a perfect expected utility maximizer, although the reverse holds true. This is because a subject may always expect certainty equivalents of the half of main prizes whatever the probability of lottery winning. In this case the subject would be classified as a risk neutral person as he/she expects, on average, the expected value of all lotteries. Subject's behavior, however, differs completely from the expected value maximization.

On the other hand being a perfectly rational decision maker guarantees the risk measure to show subject's risk neutrality. This, however, does not mean that dissimilarity from expected value maximization is a more general measure than risk attitude. Some counterexamples may be given showing that subjects presenting similar level of dissimilarity may demonstrate different risk attitudes.

8.5. The values of *RSA* may easily be calculated for each subject. The ordering of subjects is presented in table 8.1.

16	6	8	14	15	11	12	1	18	3	22	20	21	2	5	9	7	24	4	23	17	10	19	13

Table 8.1. Ordering of subjects by their Risk Seeking attitude. More risk averse subjects to the left.

[7] This simple and apparently obvious definition may, anyhow, be regarded as a very controversial. This is due to the fact it does not involve the shape of the utility function, which is the classical approach to the subject. Besides it does not take into account other more complex concepts resulting from decades long discussion on what risk aversion is. For instance Prospect Theory endeavors to separate the lottery results into the value and the probability weighting functions. However it then combines both functions in order the present what is called "the fourfold pattern of risk attitude". The author of this paper sees no reason, especially in the nonparametric approach, to split the observation into several functions, which then have to be joined in order to predict the subject's behavior. Therefore this simple definition is used further in the paper.

The values of risk seeking adjusted measure are presented in Figure 8.1.

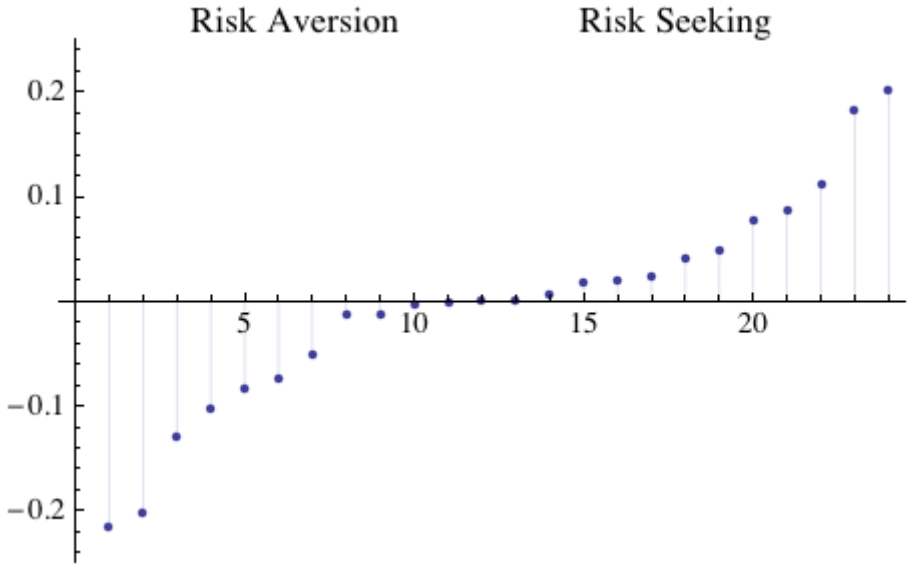


Figure 8.1. Risk Seeking adjusted measure for all subjects. Ordering as in Table 8.1.

It is seen that the most risk-averse subjects are subjects 16 and 6, whereas the most risk-seeking are subjects 13 and 19. Interestingly, the subjects, which present the average risk attitude (i.e. are middle points in Table 8.1 and in Figure 8.1) are subjects 20 and 21, so exactly those that were determined as the median subjects in the group. As their *RSA* measure is 0 it may be stated that they present a risk neutral attitude[8]. This confirms the conclusion presented in Point 7 that these subjects are perfectly rational decision makers.

9. Clustering subjects

9.1. Clustering is a method of organizing data in such a way that they are divided into separate groups (clusters) so that members of one cluster are somehow similar to each other and are dissimilar from members of other clusters. Clustering may be performed in several ways. A very basic approach has already been presented in Point 4. Detecting subjects which are the nearest to other subjects enables one to determine groups of subjects presenting similar behavior. This simplified method, however, does not consider all distances between subjects and may only serve for initial considerations.

9.2. More elaborate methods take into account all distances and use optimization algorithms. There are several algorithms known in the literature such as K-Medoid and K-Mean

[8] As stated in footnote 5 the definitions of risk seeking (7.7) and risk aversion (7.8) do not involve the shape of the utility function. It is, however, interesting to check in the Appendix that the relative utility functions of subjects 16 and 6 are generally convex, the relative utilities of subjects 13 and 19 are generally concave, and the relative utilities of subjects 20 and 21 are linear. This is in full accordance with the traditional approach of determining risk attitudes based on the shapes of utility functions.

(Dunham, 2002). These algorithms divide data into K clusters, where in the first case the cluster center is one of the data points, whereas in the second case it may be any point in between. The aim of these algorithms is to minimize the total distance between the data points. Details of the methods are not presented here as they are described in the literature and because the algorithms are available in most of the advanced statistical packages.

9.3. In this research we use the algorithm available in the Mathematica[9] program. It uses K-Medoid procedure with several distance measures as an option. The results[10] for different number of clusters and using the Euclidean Distance measure are presented in Table 9.1.

3 clusters
1 2 3 4 5 7 9 12 15 18 20 21 22 23 24
6 8 11 14 16
10 13 17 19
4 clusters
1 2 3 4 5 7 12 15 18 20 21 22 23 24
6 8 11 14 16
9
10 13 17 19
5 clusters
1 2 3 4 7 12 15 18 20 21 22 23 24
5
6 8 11 14 16
9
10 13 17 19
6 clusters
1 5 15 18 24
2 3 4 7 12 20 21 22 23
6 11 14 16
8
9
10 13 17 19

Table 9.1. Division of subjects into 3, 4, 5, and 6 clusters using K-Medoid algorithm and the Euclidean Distance measure.

Very interestingly, the division of subjects into 3 clusters resulted in a solution depending strongly on stated risk attitudes. The third cluster contains subjects 10, 13, 17, and 19, which, according to Table 8.1., present the most risk-seeking attitude. The second cluster contains subjects 6, 8, 11, 14, and 16, which present the most risk-averse attitude. The only exception is the lack in this group of subject 15, which is slightly more risk-averse than subject 11. The remaining subjects form the first cluster. As concluded in Points 4, 5, and 7 there is a large group of subjects presenting average behavior with subjects 20, 21, 22, 3, and 7 as

[9] Wolfram Research Inc.
[10] The resulting clusters may depend on starting options. Therefore it is important to use a large number of optimization steps.

its main representatives. The obtained clustering result justifies, therefore, the definitions of risk attitude presented in Point 8 and proves that risk attitude is the key factor in clustering the subjects according to their behavior in risky choices.

The results of clustering into 6 clusters are also worth discussion. The algorithm separated risk seeking subjects (10, 13, 17, and 19), risk averse subjects (6, 11, 14, and 16), risk neutral subjects whose decision making is very similar to expected value maximization (2, 3, 4, 7, 12, 20, 21, 22, and 23), and risk neutral subjects whose decision making differs from expected value maximization (1, 5, 15, 18, and 24). The remaining clusters are formed by a single subject 8, which is the outlier and by a single subject 9, which is also risk-neutral, but the decision making is highly dissimilar from the expected value maximization. The obtained results show that the dissimilarity from perfectly rational decision-making is the second key factor important during the clustering procedure[11].

9.4. The results presented in previous the subpoint enable the proposal of another simplified method of clustering the subjects. All subjects may be presented on one plane, where an x-axis is the risk attitude and y-axis is the dissimilarity from the expected value maximization. This is presented in Figure 9.1.

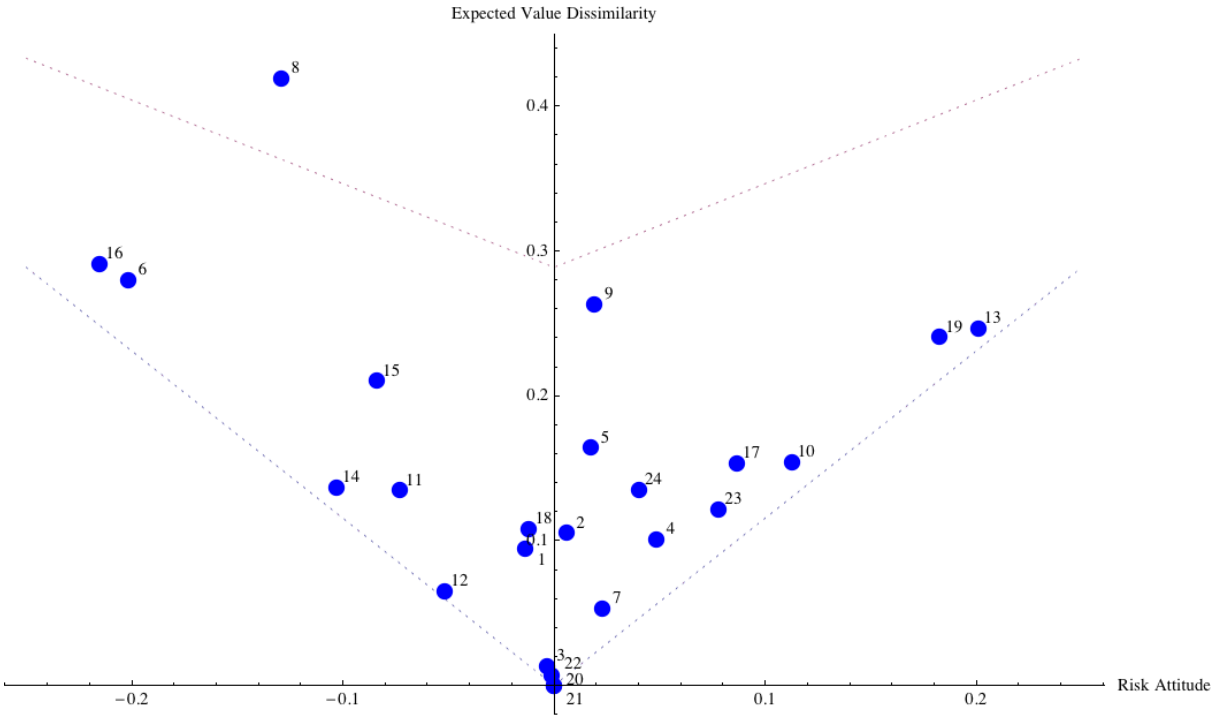


Figure 9.1. Subjects presented on Risk Attitude- Expected Value Dissimilarity Plane.

First of all it must be stated that risk attitude different from neutral also causes dissimi-

[11] Of course the K-Medoid algorithm does not know this and does not work like this. It simply minimizes the total distance. The presented conclusion results from other conclusions presented earlier in the paper.

larity from expected value maximization. Therefore the points are located only on the part of the plane. The figure presents dotted lines, which approximately limit the area of possible behaviors. The point located at (0, 0) indicates the perfectly rational maximizer. The point (0, 0.28) corresponds with the subject expecting always half of the main prize whatever the probability of winning the lottery (as in the example given in Point 8.4). The point (0.5, 0.58) – not shown in the figure - corresponds with the maximally risk seeking person, whereas point (-0.5, 0.58) corresponds with the maximally risk averse person.

These lines help to understand the behavior of respective subjects presented as dots and help to determine clusters and outliers. The results obtained by the K-Medoid algorithm and presented in Point 9.3. may now be analyzed once again[12]. Subjects 16, 6, 14, and 11 form one of the clusters as they are the most risk averse subjects in the group. Subject 8 and 15 are not included in this cluster as they are dissimilar from the expected value maximization. Subjects 13, 19, 10, and 17 form another cluster, as they are the most risk-seeking subjects in the group. Subjects 1, 5, 15, 18, and 24 are separated from other subjects having similar risk neutral attitude as they present a certain level of dissimilarity from the expected value maximization. Subject 9 as an extraordinarily case of dissimilarity is excluded and forms yet another cluster.

The example shows that the proposed plane may be a useful tool to analyze the results obtained using computer algorithms or may even serve to replace them.

10. Conclusions

This paper presented a nonparametric approach to classify data from lottery experiments. The paper demonstrated that many important questions and problems might be answered without involving any particular model. These questions concern sample "average" and extreme values, dissimilarity of subject's behavior from perfectly rational decision making and risk attitudes presented by subjects. The paper shows how subjects may be clustered in a group of similar behaviors. The only tools used for this are distance measures: Manhattan or Euclidean. Comparison of the obtained results with the relative utilities of subjects validates both methods of analyzing data from lottery experiments.

Appendix

A.1. The relative utility model assumes that certainty equivalents are transposed to the

[12] The best to compare these results also with the relative utilities presented in the Appendix.

range [0,1] using the same (3.1) transformation, which is used in this paper to normalize vectors of certainty equivalents. All the data and the estimated relative utility function are then presented on a single *p* (probability) – *r* (relative certainty equivalent) graph. No probability weighting function is needed to describe experimental results. Further details are to be found in Kontek (2010).

The Figure with the relative utilities of all subjects is presented on the next page.

The relative utility function is described using Cumulative Beta Distribution and least squares procedure was used to derive its parameters. Two outcome lotteries are presented as blue dots, more than two outcome lotteries with red dots.

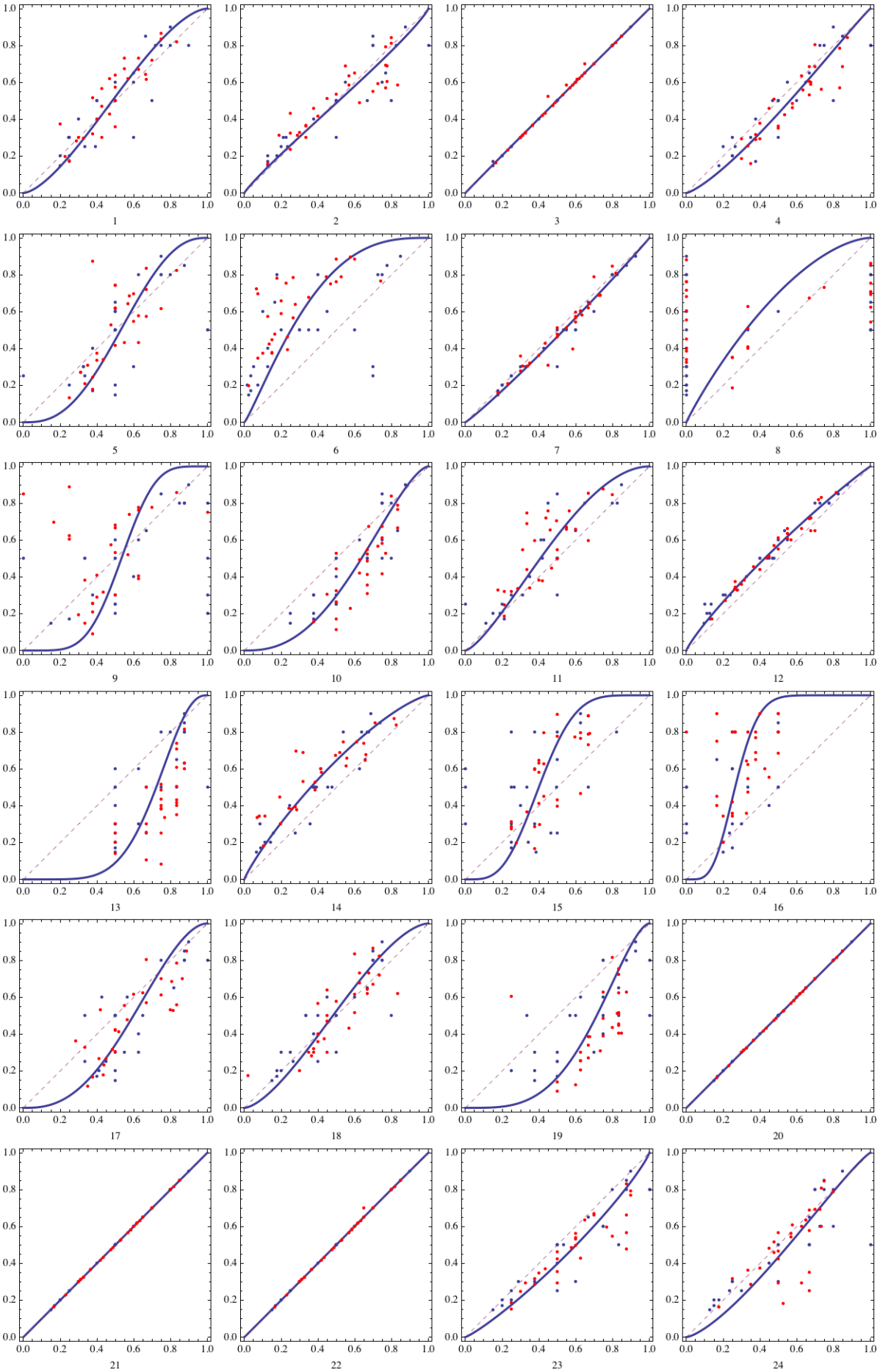
References:

1. Dunham, M., (2002). *Data Mining: Introductory and Advanced Topics*. Prentice Hall.

2. Hey, J. D., Morone, A., Schmidt, U., (2009). *Noise and bias in eliciting preferences.* J. Risk & Uncertainty, 39, pp 213-235.

3. Kontek, K., (2010). *Multi-Outcome Lotteries: Prospect Theory vs. Relative Utility.* Available at SSRN: http://ssrn.com/abstract=1617225 and as MPRA Working Paper http://mpra.ub.uni-muenchen.de/22947/

4. Traub, S., Schmidt, U., (2009). *An Experimental Investigation of the Disparity between WTA and WTP for Lotteries,* Theory & Decision, 66, pp 229-262.


1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24