A Method to Maximize the Information of a Continuous Variable in Relation to a Dichotomous Grouping Variable: Cutpoint Analysis

András Vargha

Professor
Károli Gáspár University
of the Reformed Church
in Budapest

E-mail: vargha.andras@kre.hu

Lars R. Bergman

Professor
Stockholm University

E-mail: lrb@psychology.su.se

In statistical analyses the researcher should normally use all the relevant information in the data. This argument has been used to advise against the habit of dichotomizing (approximately) continuous variables. However, if, for instance, a continuous variable is not normally distributed, it is possible that an optimal dichotomization can reveal relationships between variables otherwise obscured. Two analytical situations when this might apply were treated: *1*. The study of the relationship between an independent dichotomous grouping variable and a dependent continuous variable and *2*. the discrimination between two groups by identifying an optimal cutpoint in one or more continuous variables, treated as the predictor(s). For these purposes, cutpoint analysis (CPA) is introduced as a method for finding an optimal categorization of a continuous variable together with a computer package (ROPstat) to carry out the analysis. Three empirical examples are given that show the usefulness of CPA as compared to conventional analyses.

KEYWORDS:
Group comparison.
Best discriminating point.
Detailed comparison of distributions.

This paper takes its starting point from either of two analytical questions:

>   *Case 1*: A dichotomous grouping variable is regarded as the independent variable and it is asked how two groups differ in a continuous dependent variable.
>
>   *Case 2*: A grouping variable is regarded as the dependent variable and it is asked how well you can discriminate between the groups using information about the values of a continuous variable, which is regarded as the independent variable, that is, the predictor of group membership. Of course, in practice the continuous variable is usually not really continuous, by "continuous" we mean here a variable that is at an approximate interval scale level and taking many values, or is at least ordinal, taking many values.

Case 1 is often addressed by a two-sample t-test of the mean difference and Case 2 is frequently covered by discriminant analysis (DA) or logistic regression analysis (LRA). However, it must be assumed that the population means carry the important information about group differences in the first case and directly using the continuous variable is the best way of representing the information contained in it in the second.

We claim that sometimes in Case 1 or 2, the information value of the continuous variable is not maximized by treating it as an interval scale variable and working only with means but instead by an optimal categorization of its value range. By "information value" we mean all possible information by which one can draw conclusions from one variable to the other. Such a situation can occur in Case 1 when group membership relates to the value of the continuous dependent variable differently in various ranges of it. It might be well that there are no mean differences between groups but still the percentages belonging to one category of the categorized continuous variable vary between groups (see intervals $(-3; P_1)$, $(P_1; P_2)$, and $(P_2; 3)$ in Figure). This line of reasoning receives some support from the not infrequent finding that dichotomous variables, as compared to the corresponding continuous variables, can be surprisingly good at detecting relationships between variables (*Farrington–Loeber* [2000]). In Case 2 an appropriately categorized version of the continuous variable may be more useful to discriminate between the groups than the original variable. This can occur when there are threshold effects of the independent variable.

Two distributions with identical means but different dispersions
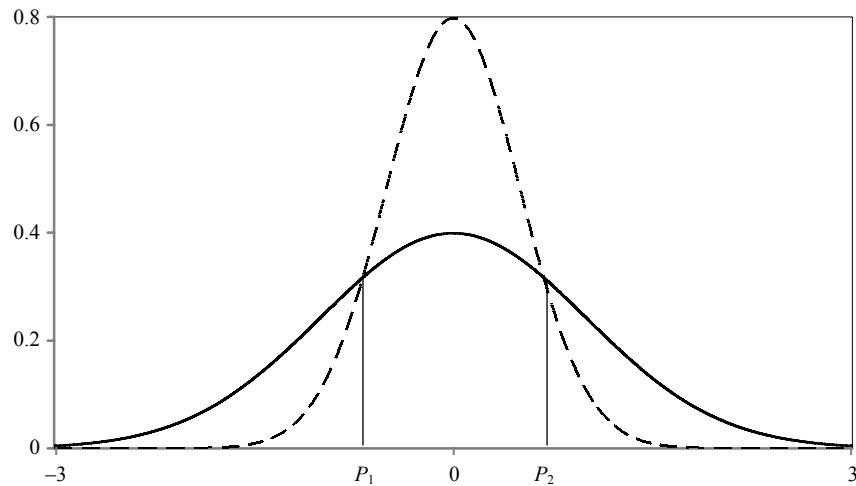
The purpose of the present paper is to present a method and a computer program to identify optimal cutpoints for categorizing a continuous variable and evaluating the usefulness of this categorization. This is done for each of the two cases described formerly and the method is called cutpoint analysis (CPA). We will only treat the case when the grouping variable is dichotomous but the findings are easily extended to the multi-group case.

First we present a brief discussion of the distributional conditions of the continuous variable in the different groups that must hold for the categorization approach to possibly detect group differences otherwise not detected or achieve a discrimination otherwise not achieved.

We note that in CPA the restrictive assumption of interval scaled continuity can be relaxed. The minimal constraint imposed is ordinality.

1. Necessary conditions for preferring an optimal categorization method

Case 1. Suppose one is interested in comparing Group A and Group B with regard to the values of a quantitative variable X. Let X be denoted in Group A by X_A, and in Group B by X_B. If X_A and X_B are normally distributed and $\sigma_A = \sigma_B$, the

only difference that can occur between the distributions of X_A and X_B is a mean difference, denoted by d:

$$F_A(x) = F_B(x + d),\qquad /1/$$

for the distribution functions F_A, and F_B for all possible values x, and for some value of d.

If the assumption of normality holds but variance homogeneity is strongly violated, large differences between the two distributions can occur even under $\mu_A = \mu_B$. (See Figure.)

If the assumption of normality is seriously violated, which occurs quite frequently (*Micceri* [1989]), then several situations can occur: Group A and Group B can differ in several special ways in terms of the distribution of X also when $\mu_A = \mu_B$. One such type of difference is when Group A members exist with a substantially lower or higher proportion under (or above) a certain value c than Group B members, that is when

$$F_A(c) < F_B(c) \ \text{ or } \ F_A(c) > F_B(c).$$

If such a cutpoint c exists on the scale of X, it suggests that different ranges of the scale of X represent different qualities. The exploration of one or more cutpoints over the range of X in relation to the distribution of X_A and X_B may then provide important information about the relationship between the independent and dependent variables.

It is also important to note that if X_A or X_B is non-normal, the inequality

$$F_A(c) \neq F_B(c)\qquad /2/$$

for some value of c can hold even if the population means of the two groups are equal, that is, $\mu_A = \mu_B$. Accordingly, the violation of normality is always an indication that differences between F_A and F_B other than a mean difference may occur.

Case 2. Whenever /2/ holds for some value c, this can serve as information for discriminating between Group A and Group B. If X_A and X_B are normally distributed, and variance homogeneity holds, then $\mu_A - \mu_B$ contains all information regarding the differences between Group A and Group B. However, if these assumptions are violated, identification of values c for which /2/ holds may suggest that there is a better rule for discriminating Group A and Group B than the one obtained in standard DA (*Farrington–Loeber* [2000]).

2. Description of the CPA method

Case 1. It is helpful if prior knowledge exists about defining the range in which the two groups differ. In a substantial portion of cases, however, the researcher has no idea about the locations of the optimal cutpoints over the range of the dependent variable X but still their existence might be surmised and merit investigation. A simple solution would be to compare the two distributions in all of the observed values of X, but this approach would necessarily yield high alpha error inflation, which is statistically unacceptable. By the CPA method presented here one can search for cutpoints discriminating sharply the two distributions without the danger of alpha inflation.

The main idea of CPA is as follows.

1. Chose a limited number of cutpoints (c_1, …, c_k) from the value range of variable X.

2. For each c_i ($i = 1$, …, k) dichotomize X at cutpoint c_i, defining variable X_i as follows: $X_i = 0$ if $X < c_i$ and $X_i = 1$ if $X \geq c_i$.

3. Compare Group A and Group B in terms of each X_i by performing a 2×2 chi-square test or a Fisher-exact test, determining the *p*-value of the significance.

4. In order to avoid alpha inflation, multiply *p* by *k*, the number of all cutpoints, that is, the number of performed tests: $p_{adj} = k \cdot p$. This is the well-known Bonferroni method for adjusting *p*-values in multiple comparisons (*Maxwell–Delaney* [2004] pp. 202–208).

The only questions left open are the choice of the value *k* and the selection of the c_i ($i = 1$, …, k) cutpoints. Whenever variable X has k_X different values, and k_X is less than k_{max}, the maximal allowed value for *k* is set to $k = k_X$. In all other cases, set $k = k_{max}$. As a rule of thumb – based on our empirical experience – we suggest that the maximal allowed value for *k* be 10 in most comparisons, that is, $k_{max} = 10$. Increasing the value of *k* would decrease the power of CPA, and a decrease of its value may increase the chance that relevant cutpoint(s) will be unidentified. In a later section we will also provide some empirical support to this choice.

In the identification process of the *k* cutpoints, we propose the following two criteria if the number of different values of variable X in the sample exceeds k_{max}:

1. Let $c_1 \geq x_\varepsilon$ and $c_k \leq x_{1-\varepsilon}$, where x_ε and $x_{1-\varepsilon}$ are percentile values in the empirical pooled distribution of variable X with certain small ε

values ($\varepsilon < 0.20$). Hence, we do not compare the groups at the lower and upper ε part of the pooled distribution. The value of ε can be fixed freely by the user – its recommended value may be between 0.01 and 0.05, depending on the sample size (in larger samples ε can be smaller, enabling CPA to detect differences between the distributions over a larger range of X values). In ROPstat the default value for ε is 0.025.

2. The estimated $P(c_i \leq X \leq c_{i+1})$ probabilities, where $1 \leq i \leq k-1$, should be as similar as possible in the total sample containing both groups.

The description of the computer program can be found in the Appendix. For more details about the program output, see the empirical examples provided in the next section.

Case 2. If the grouping variable is regarded as the dependent variable, an important aim of the analysis can be to predict group membership based on the value of the continuous variable. This model is well-known, for instance, in research evaluation and comparison of the performance of diagnostic tests (*DeLong–DeLong–Clarke-Pearson* [1988]).

The key concepts are as follows. One of the two groups is regarded as the criterion group, the other as the control group. Based on the continuous variable X, it is decided if a subject belongs to the criterion group or not. The decision is made by means of a threshold value x_c, on the scale of X. Subjects having X values greater than or equal to x_c will be regarded as belonging to the criterion group. A threshold value x_c works well if most subjects from the criterion group will be judged as belonging to it, in other words, if

$$\text{Sensitivity}\,(x_c) = Pr\big(X \geq x_c \big| \text{criterion group}\big) \qquad \text{/3/}$$

is close to 1, and most subjects from the control group will be judged as not belonging to the criterion group, that is, if

$$\text{Sensitivity}\,(x_c) = 1 - Pr\big(X \geq x_c \big| \text{control group}\big) \qquad \text{/4/}$$

is close to 1. The choice of an appropriate threshold value x_c can be made by means of a receiver operating characteristic (ROC) curve (*DeLong–DeLong–Clarke-Pearson* [1988]).

For an optimal x_c value, the $Pr(X \geq x_c)$ proportions for the criterion and control group might differ substantially. Accordingly, cutpoints identified in a CPA will

carry diagnostic information for discriminating the two groups. Starting from several continuous variables to be used for the group discrimination and then creating one or more new dichotomous variables by means of the identified cutpoints, the set of these derived variables may serve as predictor variables in a DA or LRA for arriving at an efficient discrimination.

3. Examples

Example 1. The femininity of applicants to psychology major. In an examination of admittance to the Psychology major at Eötvös Loránd University, Budapest in 1981, the number of males and females were $m = 16$, and $n = 78$, respectively. Among these 94 applicants, 12 males and 70 females filled out a short Hungarian version (including 300 items) of the California Personality Inventory (SCPI) (*Oláh* [1985]). One of the scales of SCPI is "femininity" (Fem), which informs about the feminine character and focus of the interest of the subject. In order to test the validity of this scale, we compared the male and female samples. For testing the equality of theoretical means (the sample means were 12.08 for males and 14.00 for females), the two-sample t-test ($t(80) = 2.954$, $p = 0.0041$) and the Welch test ($W(13) = 2.372$, $p = 0.0337$) were applied. For examining the stochastic equality of males and females, the Mann–Whitney test ($z = 2.339$, $p = 0.019$) and the Brunner–Munzel test ($BM(12) = 2.108$, $p = 0.0566$) were performed (about stochastic equality, see *Vargha– Delaney* [1998], [2000]). The estimated A measure of stochastic superiority, which assesses the stochastic dominance of males versus females in terms of the Fem scale was 0.29. This shows that if we compare two randomly selected male and female persons among the applicants, the chance of having a larger male Fem score is about 0.29, and the chance of having a larger female score is about 0.71.

For a detailed comparison of the two distributions, a cutpoint analysis was performed by means of the group comparison module of ROPstat. The program divides the scale of the dependent variable, X into many narrow intervals so that the cutpoints define intervals with as equal proportion of the total sample as is possible, and if the number of the different values of X does not exceed 100, each value will be placed in the inner part of a separate interval. If the number of different values of X exceeds 100, some values of X may fall to the edges of these intervals. The program computes the value of the empirical distribution function for the upper limits c of these intervals separately for the compared groups (males, $F1(c)$ and females, F2(c)), and tests their difference at $k \leq 10$ different points. In the present example,

the program found 10 different values (between 8 and 18), but for the two lowest c values (8.05 and 10.05) the pooled cumulative percentage value was less than 0.05, and for the largest (18.05) was greater than 0.95. Hence in this case $k = 7$ and $\varepsilon = 0.05$. Based on the F1(c) and F2(c) values corresponding to the selected 7 cutpoints, the program computes for each c the φ contingency-coefficient measuring the strength of association between the grouping variable (in the present case "gender") and the dichotomized dependent variable (in the present case femininity), using either the 2x2 chi-square test or the Fisher-exact test with the corresponding unadjusted and adjusted two-tailed probability values. Due to the relatively small sample sizes in the present case, the Fisher-exact test was always performed. The results are summarized in Table 1.

Table 1

Results from a cutpoint analysis comparing males and females based
on their California Personality Inventory/femininity level
($n = 82$)

Detailed point-wise comparison of the two distribution functions							
c	$F1(c)$	$F2(c)$	$F1–F2$	Phi	Chi|Fish	p-value	Adjusted p
8.05	0.083	0	0.083	0.27			
10.05	**0.417**	**0.029**	**0.388**	**0.49**	**Fisher**	**0.0005**	**0.0036***
11.05	0.500	0.100	0.400	0.39	Fisher	0.0027	0.0191**
12.05	0.583	0.229	0.355	0.28	Fisher	0.0314	0.2196
13.05	0.667	0.414	0.252	0.18	Fisher	0.1259	0.8814
14.05	0.750	0.629	0.121	0.09	Fisher	0.5250	1.0000
15.05	0.833	0.757	0.076	0.06	Fisher	0.7232	1.0000
16.05	1.000	0.900	0.100	0.13	Fisher	0.5861	1.0000
17.05	1.000	0.943	0.057	0.09			
18.05	1.000	1.000					

Note. The significance of phi is tested via chi-square or Fisher-exact test. Tail probability (p) is adjusted by means of Bonferroni method. * for $p < 0.05$ ** stands for $p < 0.01$;

To analyze the identity of the two distributions, Kolmogorov–Smirnov's two-sample test is applied: $J* = 1.280$ ($p = 0.0754$)

The format of the table follows that of the corresponding computer output of ROPstat with some modifications.

Based on the results summarized in Table 1, the most significant difference between the two genders was obtained for the cutpoint $c = 10.05$ (the corresponding row in Table 1 is indicated by a bold type face). Below this value (that is in the range

0–10) we find 41.7 percent of males and 2.9 percent of females. The difference is 39 percent, which is highly significant (adjusted $p = 7 \cdot 0.0005 \approx 0.0036$). It is interesting that the Kolmogorov–Smirnov test shows only a tendency to significance ($p = 0.0754$). This weakness of the test is characteristic and is due to the fact that it performs a global comparison of two distributions, taking into account every possible type of difference, whereas CPA focuses on a small number of potentially informative cutpoints.

The psychological explanation of the obtained results may be as follows. The femininity of males and females differs from each other mostly in the fact that there exists a certain level of minimal femininity ($\text{Fem} = 10$), below which we find almost exclusively males. Among the females almost everybody (in the present sample 68 out of 70) shows this minimal level of femininity. Such a strong differentiation, however, does not occur at the higher region of the Fem scale, which means that there is not a high level of femininity that would mainly be characteristic of females. This information is not revealed by standard analyses.

Example 2. The relationship of birth rank to personality. This study concerns the relationship of birth rank to adult personality (*Mózes–Vargha* [2007]). Studying women, we compared first born subjects ($m = 35$) with the rest of the sample ($n = 49$) in terms of six scales of Parker's parental bonding instrument (*Parker* [1989], [1990]). Among these six scales we present results concerning father's care. The measure is "retrospective", meaning that the women reported how they remember their father cared for them during their first 16 years.

Table 2

Comparison of first born and other women's responses to the father's care scale
in the parental bonding instrument
($n = 84$)

Group	Size	Mean	SD	Minimum	Maximum	Skewness	Kurtosis
First born women	35	23.17	11.44	0	36	–0.829*	–0.428
Other women	49	21.96	8.075	3	34	–0.651+	–0.508

Note. Dependent variable is father's care. The significant skewness indicated non-normality. + stands for $p < 0.10$; * for $p < 0.05$.

Testing the equality of population variances: *1*. O'Brien test (Welch type): $F(1.0; 45.6) = 4.851$ ($p = 0.0327$); *2*. Levene test (Welch type): $F(1; 55.4) = 3.645$ ($p = 0.0614$).

Testing the equality of population means: *1*. Two-sample *t* test: $t(82) = 0.570$ ($p = 0.5704$); *2*. Welch's modified t test: $W(57.4) = 0.538$ ($p = 0.5924$).

The format of the table follows that of the corresponding computer output of ROPstat with some modifications.

For testing the equality of theoretical means, two-sample *t*-tests were applied but none of them indicated any significant difference between the two groups. (See Table 2.) Nonparametric rank tests comparing the two groups were also far from being significant ($p > 0.20$). However, in the present case, both the normality assumption and the variance homogeneity assumption are violated. On the one hand, this may invalidate the two-sample *t*-tests, on the other hand, it raises the possibility that some other types of differences may appear using CPA. The results of this analysis are summarized in Table 3.

Table 3

Results from CPA comparing first born and other women in terms
of the father's care scale in the parental bonding instrument
($n = 84$)

Detailed point-wise comparison of the two distribution functions							
c	*F*1(*c*)	*F*2(*c*)	*F*1–*F*2	Phi	Chi|Fish	*p*-value	Adjusted *p*
0.18	0.057	0.000	0.057	0.18			
1.26	0.086	0.000	0.086	0.23			
3.06	0.114	0.020	0.094	0.20			
5.22	0.143	0.041	0.102	0.18	Fisher	0.1223	1.0000
6.30	0.171	0.041	0.131	0.22			
7.02	0.171	0.061	0.110	0.18			
8.10	0.171	0.082	0.090	0.14			
9.18	0.171	0.102	0.069	0.10			
10.26	0.171	0.122	0.049	0.07			
11.34	0.171	0.143	0.029	0.04			
12.06	0.171	0.163	0.008	0.01			
13.14	0.200	0.184	0.016	0.02	Fisher	1.0000	1.0000
14.22	0.229	0.224	0.004	0.00			
16.02	0.229	0.245	–0.016	–0.02			
18.18	0.257	0.306	–0.049	–0.05	Fisher	0.8069	1.0000
19.26	0.286	0.347	–0.061	–0.06			
20.34	0.314	0.367	–0.053	–0.06			
21.06	0.343	0.388	–0.045	–0.05	Fisher	0.8191	1.0000
22.14	0.371	0.449	–0.078	–0.08			
23.22	0.429	0.510	–0.082	–0.08	Fisher	0.5112	1.0000
24.30	0.486	0.510	–0.024	–0.02			
25.02	0.514	0.612	–0.098	–0.10	Fisher	0.3826	1.0000
26.10	0.514	0.633	–0.118	–0.12			

(Continued on the next page.)

(Continuation.)

			Detailed point-wise comparison of the two distribution functions				
c	$F1(c)$	$F2(c)$	$F1–F2$	Phi	Chi|Fish	p-value	Adjusted p
27.18	0.600	0,673	–0.073	–0.08	Fisher	0.4993	1.0000
28.26	0.629	0.776	–0.147	–0.16	Fisher	0.1522	1.0000
29.34	0.629	0.816	–0.188	–0.21			
30.06	0.629	0.857	–0.229	–0.26	Fisher	0.0202	0.2017
31.14	0.657	0.939	–0.282	–0.36			
32.22	0.714	0.959	–0.245	–0.35			
33.30	**0.714**	**0.980**	**–0.265**	**–0.39**	**Fisher**	**0.0005**	**0.0051****
34.02	0.914	1.000	–0.086	–0.23			
36.18	1.000	1.000					

Note. To test the identity of the two distributions, Kolmogorov-Smirnov's two-sample test was applied: $\left(J^* = 1.273\right)$ ($p = 0.0784$).

The format of the table follows that of the corresponding computer output of ROPstat with some modifications.

$F1$ refers to the distribution function for first born women and $F2$ to the corresponding function for other women. + stands for $p < 0.10$; * for $p < 0.05$, and ** for $p < 0.01$.

In our case the best discriminating point is $c = 33.30$ (the corresponding row in Table 3 is indicated by a bold type face). A lower value, that is, $X \le 33$ occurred for 71.4 percent (25 out of 35) of first born women, and 98 percent (48 out of 49) of other women. These two proportions differ from each other significantly (the two-tailed probability of the Fisher-exact test is $p = 0.0005$). This is highly significant even after performing the Bonferroni adjustment ($p_{adj} = 0.005$). Accordingly, we can claim that first born women are significantly more likely (in the present sample the chance is 28.6 percent) to report an extreme high level ($X > 33$) of father's care as compared to other women (in this latter sample the chance is 2 percent). The conclusion is that a very high level of experienced father's care is almost only found among first born women.

Example 3. Discrimination of psychotic and normal women by means of psychiatric rating scales. In the framework of a longitudinal study launched in 1967, 230 psychotic and 41 mentally normal women were investigated by means of *Overall's* [1968] factor construct rating scale (FCRS) and *Rockland* and *Pollin's* [1965] questionnaire (RPQ) for psychiatric rating (*Pethő* [2001]). In the current analysis we used 17 elementary scales of FCRS ($F1$, ..., $F17$) and 34 elementary scales of RPQ ($R1$, ..., $R33$, $R35$). A value of zero on these scales reflects the lack of some psychiatric symptom, and values close to the maximum show the strong presence of a symptom. Preliminary analyses indicated that several of the scales were non-normally distrib-

uted. For these variables we addressed the following question concerning the discrimination of psychotic and normal subjects: If we dichotomize the continuous variables, using cutpoints identified by a CPA, and then perform DA and LRA, will we arrive at a better group discrimination as compared to conventional analyses based on the continuous variables? The following statistical analyses were undertaken:

1. First a CPA was carried out for all continuous variables. In the subsequent analyses we retained only those for which the CPA revealed at least one significant cutpoint. For these scales we performed a dichotomization at the cutpoint that had the lowest *p*-value. These scales were as follows: $F1$–$F14$, $F16$, $F17$, $R1$–$R5$, $R9$, $R11$–$R16$, $R18$, $R20$–$R23$, $R25$–$R30$, and $R33$, altogether 40 variables.

2. Subsequently, we performed stepwise DA and LRA with first the original 51 continuous variables, then with the 40 dichotomized variables to predict group membership. The results are summarized in Table 4.

Table 4

Percentage of correct identifications in stepwise discriminant analyses
and binary logistic regression analyses for the factor construct rating scale
and Rockland and Pollin's questionnaire scales in original and dichotomized form

Group	Discriminant analysis with original variables	Discriminant analysis with dichotomized variables	Logistic regression analyses with original variables	Logistic regression analyses with dichotomized variables
Psychotic ($n = 230$) (percent)	78.7	87.0	95.7	95.7
Normal ($n = 41$) (percent)	92.7	100.0	82.9	85.4
Total (percent)	80.8	88.9	93.7	94.1
Number of selected variables	11	7	10	8

Based on Table 4 we can draw the following conclusions.

1. Using non-normal independent variables in DA may lead to substantially weaker discrimination than LRA.

2. Using derived dichotomized variables may lead to surprisingly good results parallel to those found using the original variables, and CPA can be an efficient tool for identifying appropriate cutpoints for the dichotomization. As an example, we obtained 88.9 percent correct identification percentage in DA with 7 selected dichotomous variables, compared to 80.8 percent with 11 original variables.

3. Also in LRA the dichotomized variables performed well (with 8 variables resulting in 94.1 percent correct classifications, as likened to using 10 original variables resulting in 93.7 percent correct classifications).

4. Summary and conclusion

It is an almost trivial observation that in statistical analyses the researcher should normally use all the relevant information in the data. In the literature there are many arguments against the habit of dichotomizing continuous variables, which is usually performed for the purpose of simplifying the analyses and presentation or for handling interactions. This attitude is seen in its most extreme form in an editorial in the *Journal of Consumer Research*, entitled "Death to Dichotomizing" (*Fitzsimons* [2008]).

The warnings against dichotomization are often good advice but the arguments build on assumptions of normality and linearity (for example *Cohen* [1983], *Maxwell–Delaney* [1993]). If these assumptions are valid, the argument against dichotomization seems solid, however, frequently psychological variables do not follow the normal distribution (*Micceri* [1989]) and the relationships might not be linear. In such situations it is possible that the arguments against dichotomization of a continuous variable break down. For instance, take the case of studying the relationship between one continuous independent variable, regarded as the risk factor, and one continuous dependent variable, regarded as the outcome. Theoretically, it is possible that there is a threshold effect in the independent variable so that there is no risk increase for a bad outcome below a certain level of the value in the risk factor but then, suddenly, a strong increase in the risk occurs. Or it is possible that the outcome is really generated by a normal mixture model with a relationship between the risk factor and the distribution membership.

Within the context discussed formerly, we examined two analytical situations where dichotomization may be appropriate. The first one concerned the study of the relationship between a dichotomous grouping variable, regarded as the independent variable, and a continuous variable, considered as the dependent variable. We devised a dichotomization method, cutpoint analysis, in which a limited number of cutpoints (usually not exceeding 10) in the dependent variable distribution are used for different dichotomizations, selecting the one that maximizes the association between the independent variable and the dependent variable. In two empirical examples, one concerning gender and femininity and the other regarding birth rank and personality, the results indicated significant relationships not revealed by standard analyses.

These findings were partly explained by clear departure from normality in the dependent variable and by the fact that the relationship had a different form in different regions of the variables.

The second analytical situation that we discussed and where dichotomization may be appropriate concerned the discrimination between two groups by identifying an optimal cutpoint in one or more continuous variables, treated as the predictor(s). CPA can then be used to find an optimal dichotomization of the continuous variable(s) in the sense that the prediction of group membership is maximized. In a third empirical example, CPA was used for dichotomizing a number of psychiatric rating scales that were used in DA or LRA. This resulted in a higher or at least as high discrimination power between psychotic and normal women as was achieved using the original continuous variables. It appears that, for discrimination purposes, the essential information in the scales was largely binary, of qualitative nature.

DeCoster, Iselin, and *Gallucci* [2009] revealed also several situations in which the use of dichotomization is appropriate. Specifically, they argue that it is acceptable for researchers to use dichotomized indicators in the following circumstances:

1. The study uses extreme group analysis.

2. The purpose of the research is to investigate how a dichotomized measure will perform in the field.

3. The underlying variable is naturally categorical, the observed measure has high reliability, and the relative group sizes of the dichotomized indicator match those of the underlying variable.

CPA is similar to the search for an "optimal cutpoint" in biostatistics. A common aim in biomedical research is to investigate whether a certain continuous variable (regarded as covariate) has potentially prognostic relevance for a time outcome dependent variable like survival. The optimal cutpoint is that value of the covariate, which corresponds to the most significant relationship between the covariate dichotomized at this cutpoint and the dichotomous outcome variable (*Heinzl–Tempfer* [2001]). Since the term "optimal" may give the false impression that this method is superior to other ones, *Altman, Lausen, Sauerbrei*, and *Schumacher* [1994] suggested that the method be called the "minimum *P*-value approach". When using this method, some researchers are ready to ignore the multiple testing and alpha inflation problems since their decisions are based on the unadjusted *p*-value of the "optimal" cutpoint. This practice may lead to inconsistent results in medical prognostic research (*Heinzl–Tempfer* [2001]). In contrast, CPA is protected against alpha inflation by performing only a limited number of two-group comparison tests (≤ 10) of scattered cutpoints and by applying a Bonferroni adjustment to the *p*-values of the selected cutpoints.

For comparing two survival curves, the most common statistical test is the logrank test. This is a type of chi-square test, asymptotically equivalent to the likelihood ratio test, which is based on observed and expected frequencies of a certain time event belonging to different time points of the two survival curves (*Bland–Altman* [2004], *Mantel* [1966], *Schoenfeld* [1981]). The need for applying some adjustment on the *p*-values of repeated logrank tests is now increasingly recognized (*Altman et al.* [1994], *Heinzl–Tempfer* [2001], *Williams et al.* [2006]). Such an adjustment can be achieved by the following formula for an adjustment of the minimal *P*-value (p_{min}) valid for large sample sizes, to allow for the multiple testing thanks to Lausen and Schumacher ([1992], [1996]), *Miller* and *Siegmund* [1982], and *Heinzl* [2000]:

$$P_{cor} = \varphi(z)\left(z - \frac{1}{z}\right)ln\left(\frac{(1-\varepsilon)^2}{\varepsilon^2}\right) + \frac{4\varphi(z)}{z}. \qquad\qquad /5/$$

Here P_{cor} denotes the adjusted (corrected) minimum *P*-value of the logrank statistic, φ is the standard normal density, z is the $\left[1-\left(P_{min}/2\right)\right]$ -quantile of the standard normal distribution and ε is defined as follows. The minimum *P*-value approach requires the choice of a selection interval. It is defined by the ε and $(1-\varepsilon)$ -quantile of the observed values of the continuous covariate $(0 < \varepsilon < 0.5)$. Values outside the selection interval are not considered as potential cutpoints. In CPA we also seek potential cutpoints between C_{ε} and $C_{1-\varepsilon}$ percentiles, so the meaning of ε in CPA is the same.

By means of formula /5/, we compared the power of the above adjustment rule for $\varepsilon = 0.01$, 0.05 and 0.10 with CPA for a set of different nominal *p*-values, allowing for as many as 10 cutpoints ($k = 10$) in CPA. Results summarized in Table 5 show that CPA is much more efficient in detecting possible differences of the two distributions to be compared than the one defined by formula /5/. This is reflected by the fact that the P_{cor} values are substantially higher – in most cases more than twice as large – than the corresponding P_{adj} values of CPA for unadjusted alpha values less than 0.05. Due to this, the cutpoints of CPA can be more easily significant despite their strictly controlled Type I error level. However, we agree with Heinzl and *Tempfer* [2001] that without any biological (clinical, psychological, etc.) indications for the actual existence of a cutpoint even the correct application of the minimum *P*-value approach as well as CPA has to be considered methodologically questionable.

Table 5

Comparison of three corrected p-values and the adjusted p-value of CPA based
on the Bonferroni method for different unadjusted nominal p-values

Unadjusted *p*-value	$P_{cor}\left(\varepsilon=0.01\right)$	$P_{cor}\left(\varepsilon=0.05\right)$	$P_{cor}\left(\varepsilon=0.10\right)$	$P_{cor}\left(k=10\right)$
0.10	1.0000	0.8806	0.7208	1.0000
0.05	0.8980	0.6183	0.4916	0.5000
0.01	0.3132	0.2087	0.1615	0.1000
0.005	0.1859	0.1231	0.0946	0.0500
0.001	0.0509	0.0334	0.0255	0.0100
0.0005	0.0285	0.0186	0.0142	0.0050
0.0001	0.0071	0.0046	0.0035	0.0010

Note. *P*-values are based on based on formula /5/ (P_{cor} with $\varepsilon=0.01$, 0.05, and 0.10).

A special type of the minimum *P*-value approach is the following method. Two groups are compared by means of a quantitative variable the same way as in CPA, looking for an "optimal" cutpoint. In this approach a cutpoint is regarded as optimal if the usual chi-square statistic computed from a 2×2 table based on the frequencies below and above the cutpoint in the two groups is maximal. Miller and Siegmund [1982] investigated the asymptotic distribution of this maximally selected chi-square statistic and provided tail probabilities and critical values for its significance for different nominal alpha levels and selection interval defined by ε and $\left(1-\varepsilon\right)$ the same way as mentioned formerly (see Tables 1 and 2 in *Miller–Siegmund* [1982]). Since the computation of these values is built on the same formula (see formula /8/ in *Miller–Siegmund* [1982] that appears in /5/), the superiority of CPA over this approach in terms of power still remains. With the same method, *Koziol* [1991] provided better critical values and tail probabilities based on the exact finite-sample distribution theory, *Betensky* and *Rabinowitz* [1999] generalized the asymptotic distribution of the maximally selected chi-square statistic for the multi-group case, and *Boulesteix* [2006] generalized the results of Koziol to any ordinally scaled dependent variable in the two-group-comparison case.

The P_{cor} corrected *p*-values of the minimum *P*-value approach refer to the significance of the most significant cutpoint that discriminates the two groups based on the continuous dependent variable, whereas the P_{adj} adjusted *p*-values of the CPA approach refer to the significance of all *k* cutpoints identified in CPA. As we could see formerly, in the practically relevant cases, when the corrected/adjusted *p* values are close to significance (this is true when the unadjusted *p*-values are less than or

equal to 0.01; see Table 5), the adjusted p-values of CPA are always substantially smaller than those of the corrected p-values of the minimum P-value approach and for this reason they can detect differences between the two distributions to be compared with a higher efficiency. This is completely true for dependent variables where the number of different values do not exceed 10 (in this case CPA compares the two distributions with all possible cutpoints). However, if the dependent variable is really continuous and has large many different values, it may well happen that the selected set of cutpoints in CPA does not include the value of the dependent variable which discriminates most significantly the two groups (distributions). To have some information about how often this unlucky situation may arise, we carried out the following empirical investigation.

From an archival data set including 811 Rorschach-protocols that served as the basis for the construction of the Hungarian Rorschach Standard (*Vargha* [1989]), we selected 236 quantitative variables of elementary Rorschach scores and computed indices. The elementary scores were relative frequencies of different Rorschach responses (referring to the location, determinant, content category, popularity, or originality of the response, etc.). These relative frequencies were computed by dividing the number of occurrences of different Rorschach-items with the number of total response number. As an example, the value of the Anat% Rorschach-variable was obtained for a specific person by dividing the number of anatomical responses occurring in the protocol by the total number of responses. More than 70 percent of these Rorschach-variables was practically continuous, having more than 10 different values. Out of the 811 protocols, 363 originated from mentally normal persons (MN), while the other 448 from institutionalized non-psychotic patients (INP).

In order to assess merits and weaknesses of CPA, the two groups (MN and INP) were compared for each of the 236 Rorschach-variables:

1. performing a CPA (with parameters $k = 10$ and $\varepsilon = 0.01$);

2. identifying the best discriminating point, that is the cutpoint within the middle $(\varepsilon; 1 - \varepsilon)$ part of the scale of the dependent variable for which the tail probability of a 2×2 chi-square test (or the two-sided p-value of the Fisher-exact test if the minimal expected cell frequency in the 2×2 table does not exceed 20) is the smallest;

3. performing the Kolmogorov–Smirnov two-sample test for a global comparison of the two distributions.

The results of the performed analyses can be summarized as follows.

4. For 157 out of the 236 dependent variables there were more then 10 different values within the middle $(\varepsilon; 1 - \varepsilon)$ part of the scale of the dependent variable. Out of these 157 variables, the smallest adjusted

P-value was less than or equal to 0.10 for 68 variables, and in 53 cases (78%) out of the 68 variables the mostly significant cutpoint was identical with the CPA cutpoint for which the tail probability of the 2×2 chi-square test was minimal. This means that the set of the *k* cutpoints of CPA contained the best discriminating point of the dependent variable in the large majority of cases. However, if one wants to decrease the risk of missing a relevant cutpoint, the value of *k* can be increased even above 10. Based on data of Table 5, one can conclude that even an increase by 50 per cent can keep the advantage of the CPA method over the alternative methods (for example maximizing chi-square).

5. For a comparison of the efficiencies of CPA and the Kolmogorov–Smirnov two-sample test, we cross-tabulated the significances of the Kolmogorov–Smirnov test and the most significant cutpoint (adjusted probability) of CPA for the 236 Rorschach-variables. (See Table 6.) From Table 6 it seems to be evident that CPA highly outperforms the Kolmogorov–Smirnov in terms of power. It occurred only four times out of 236 cases that the Kolmogorov–Smirnov test was significant (3 times at 10 per cent and once at 5 per cent level), but the CPA was not, whereas the opposite situation occurred in 38 cases. In addition, in 31 (out of the 38) cases, the CPA was significant at least two levels stronger than the Kolmogorov–Smirnov test (the opposite situation never occurred)

Table 6

Cross-tabulation of the significances of the Kolmogorov–Smirnov test and the most significant cutpoint
of the cutpoint analysis for 236 different quantitative Rorschach variables

Kolmogorov–Smirnov test	CP					
	$p > 0.10$	$p < 0.10$	$p < 0.05$	$p < 0.01$	$p < 0.001$	*Total*
$p > 0.10$	137	18	12	5	3	*175*
$p < 0.10$	3	7	2	5	0	*17*
$p < 0.05$	0	1	6	7	6	*20*
$p < 0.01$	0	0	0	6	5	*11*
$p < 0.001$	0	0	0	0	13	*13*
Total	*140*	*26*	*20*	*23*	*27*	*236*

CPA can be regarded as a multiple test like post-hoc analyses in ANOVA. The *k* cutpoints are selected based on the distributional characteristics of the pooled sam-

ple, independently from the differences between the two groups. The application of the well-known Bonferroni method guaranties that if any of the cutpoints is significant at an adjusted alpha level, then the probability of Type I error (of the null hypothesis of the equality of the two distributions binarized in this cutpoint) will not exceed α. Hence, if any cutpoint of CPA is significant, the two distributions can confidently be declared different from each other. Since CPA gave significant results in many more cases than the Kolmogorov–Smirnov test, it seems in this context to be more appropriate for detecting differences. It is also important to add that CPA not only detects efficiently the inequality of the two distributions but identifies also the cutpoints where the differences are most salient.

It should be noted that if the same data set is used both to dichotomize variables by CPA and to estimate a regression model with the dichotomized variables as predictors, the corresponding regression coefficients will obviously be biased. In such cases we suggest that, if the sample size is large enough, a portion (say 2/3 of the sample) be used for exploration and the rest for the verification of the model. If the sample is relatively small, the results of CPA need confirmation in an independent study. We assert, however, that, at a minimum, CPA is a simple but effective way of deriving hypotheses to be confirmed in future studies.

To sum up, CPA is a new technique and software for finding efficiently truly significant dichotomizing points in a quantitative variable that maximizes the association to another dichotomous variable, which might otherwise be hidden if one were to use conventional statistical approaches.

In the present article, only two special cases were treated but we believe that the reasoning employed merits consideration also in other ones. The most obvious extensions are to the case where the grouping variable is not dichotomous and to the one where the relationship is studied between two variables while controlling for a third. ROPstat can handle the multigroup case of CPA whose nice empirical illustration can be found in *Borbély–Vargha* [2010].

Appendix

The description of the computer program

We implemented CPA in the group comparison module of ROPstat, a new statistical program package. It is a user-friendly statistical software that is rich in robust techniques and procedures with ordinally scaled variables, and includes a number of procedures for pattern and person oriented analysis.

Its free demo version can be downloaded from the site www.ropstat.com by clicking on the text "Download and test DEMO version in English". A description of the package can also be found there. A CPA is carried out in the following way.

1. Run the downloaded setup program of ROPstat. As a result, a folder called "c:_vargha\ropstat" will be created and a program called "ropstat.exe" will be installed in it.

2. Run ropstat.exe.

3. Open an input data file by means of icon "Open". ROPstat has a special option accepting SPSS portable files imported from SPSS in *.por format and tab-delimited files imported from Excel in *.txt format. The DEMO version of ROPstat accepts at most 5 variables and 500 cases but otherwise performs complete statistical analyses.

4. After loading a data file, click on the menu point "Statistical analyses", and within it the submenu points "Comparing groups or variables" and "One-way comparison of independent samples".

5. In the appearing program window, put the given continuous variable (X) from the list of variables to the box of "Dependent variables", and a grouping variable having two code values (or defined by two intervals in the variable characteristics window of the data sheet) to the box of "Grouping variable".

6. In the box of "Scale type"', change the scale type of the dependent variable from "interval" to "ordinal", and change the option of "Detailed comparison of distributions" in this program window from "No" to "Yes".

7. When you click on the icon "Run" in the bottom of the program window, a list of the following results will appear in a text window:

a) Nonparametric group comparison with the classical Mann–Whitney test.

b) Two robust alternatives of the Mann–Whitney test (Brunner–Munzel and corrected Fligner–Policello tests).

c) Detailed point-wise comparison of the two distributions (CPA). The two groups are compared by a 2×2 chi-square test if the minimal expected cell frequency exceeds 20, otherwise by the Fisher-exact test. Here $k = min\left(10, \text{ number of different values of } X\right)$.

d) Kolmogorov–Smirnov's two-sample test for a global comparison of the two distributions.

8. In the CPA part of the output, the rightmost column with the header "Adjusted p" will contain the adjusted *p* values for the point-wise comparisons. If such a *p* value is less than 0.05, the corresponding score of the test variable can be regarded as a significant cutpoint.

References

ALTMAN, D. G. – LAUSEN, B. – SAUERBREI, W. – SCHUMACHER, M. [1994]: Dangers of Using "Optimal" Cutpoints in the Evaluation of Prognostic Factors. *Journal of the National Cancer Institute*. Vol. 86. No. 11. pp. 829–835.

Betensky, R. A. – Rabinowitz, D. [1999]: Maximally Selected Chi-Square Statistics for kx2 Tables. *Biometrics*. Vol. 55. No. 1. pp. 317–320.

Bland, J. M. – Altman, D. G. [2004]: The Logrank Test. *British Medical Journal*. Vol. 328. No. 7447. p. 1073.

Borbély, A. – Vargha, A. [2010]: Az *l* variabilitása öt foglalkozási csoportban – Kutatások a Budapesti Szociolingvisztikai Interjú beszélt nyelvi korpuszban. *Magyar Nyelv*. Vol. 106. No. 4. pp. 455–470.

Boulesteix, A. L. [2006]: Maximally Selected Chi-Square Statistics for Ordinal Variables. *Biometrical Journal*. Vol. 48. No. 3. pp. 451–462.

Cohen, J. [1983]: The Cost of Dichotomization. *Applied Psychological Measurement*. Vol. 7. No. 3. pp. 249–253.

DeCoster, J. – Iselin, A-M. R. – Gallucci, M. [2009]: A Conceptual and Empirical Examination of Justifications for Dichotomization. *Psychological Methods*. Vol. 14. No. 4. pp. 349–366.

DeLong, E. R. – DeLong, D. M. – Clarke-Pearson, D. L. [1988]: Comparing the Area Under Two or More Correlated Receiver Operating Characteristic Curves: A Nonparametric Approach. *Biometrics*. Vol. 44. No. 9. pp. 837–845.

Farrington, D. P. – Loeber, R. [2000]: Some Benefits of Dichotomization in Psychiatric and Criminological Research. *Criminal Behaviour and Mental Health*. Vol. 10. No. 2. pp. 100–122.

Fitzsimons, G. J. [2008]: Death to Dichotomizing. *Journal of Consumer Research*. Vol. 35. No. 1. pp. 5–8.

Heinzl, H. [2000]: Dangers of Using "Optimal" Cutpoints in the Evaluation of Cyclical Prognostic Factors. In: *Ferligoj, A. – Mrvar, A.* (eds): *New Approaches in Applied Statistics*. Metodološki zvezki, 16. FDV. Ljubljana.

Heinzl, H, – Tempfer, C. [2001]: A Cautionary Note on Segmenting a Cyclical Covariate by Minimum P-Value Search. *Computational Statistics & Data Analysis*. Vol. 35. Issue 4. pp. 451–461.

Koziol, J. A. [1991]: On Maximally Selected Chi-Square Statistics. *Biometrics*. Vol. 47. No. 4. pp. 1557–1561.

Lausen, B. – Schumacher, M. [1992]: Maximally Selected Rank Statistics. *Biometrics*. Vol. 48. No. 3. pp. 73–85.

Lausen, B. – Schumacher, M. [1996]: Evaluating the Effect of Optimized Cutoff Values in the Assessment of Prognostic Factors. *Computational Statistics & Data Analysis*. Vol. 21. Issue 3. pp. 307–326.

Mantel, N. [1966]: Evaluation of Survival Data and Two New Rank Order Statistics Arising in Its Consideration. *Cancer Chemotherapy Reports*. Vol. 50. No. 3. pp. 163–170.

Maxwell, S. E. – Delaney, H. D. [1993]: Bivariate Median Splits and Spurious Statistical Significance. *Psychological Bulletin*. Vol. 113. No. 1. pp. 181–190.

Maxwell, S. E. – Delaney, H. D. [2004]: *Designing Experiments and Analyzing Data: A Model Comparison Perspective*. Lawrence Erlbaum Associates. Mahwah.

Micceri, T. [1989]: The Unicorn, the Normal Curve, and Other Improbable Creatures. *Psychological Bulletin*. Vol. 105. No. 1. pp. 156–166.

Miller, R. – Siegmund, D. [1982]: Maximally Selected Chi Square Statistics. *Biometrics*. Vol. 38. No. 4. pp. 1011–1016.

MÓZES, T. – VARGHA, A. [2007]: A születési sorrend és a személyiség összefüggései. In: *Bagdy*, *E.* *– Mirnics*, *Zs*. *– Vargha*, *A*. (eds.): *Egyén–Pár–Család. Tanulmányok a pszichodiagnosztikai tesztadaptációs és tesztfejlesztő kutatások köréből*. pp. 249–270. Animula. Budapest.

OLÁH, A. [1985]: A Californiai Pszichológiai Kérdőív hazai alkalmazásával kapcsolatos tapasztalatok. In: *Hunyady*, *G*. (ed.): *Pszichológiai Tanulmányok. XVI.* pp. 53–101.

OVERALL, J. E. [1968]: Standard Psychiatric Symptom Description: The Factor Construct Rating Scale (FCRS). *Triangle*: *Sandoz Journal of Medical Sciences*. Vol. 8. No. 5. pp. 178–186.

PARKER, G. [1989]: The Parental Bonding Instrument: Psychometric Properties Reviewed. *Psychiatric Developments*. Vol. 7. No. 4. pp. 317–335.

PARKER, G. [1990]: The Parental Bonding Instrument: A Decade of Research. *Social Psychiatry and Psychiatric Epidemiology*. Vol. 25. No. 6. pp. 281–282.

PETHŐ, B. [2001]: *Klassifikation, Verlauf und Residuale Dimension der Endogenen Psychosen*. Platon Verlag Budapest. Universitätsverlag. Ulm.

ROCKLAND, I. H. – POLLIN, W. [1965]: Quantification of Psychiatric Mental Status. *Archives of General Psychiatry*. Vol. 12. No. 1. pp. 23–28.

SCHOENFELD, D. [1981]: The Asymptotic Properties of Nonparametric Tests for Comparing Survival Distributions. *Biometrika*. Vol. 68. No. 1. pp. 316–319.

VARGHA, A. [1989]: *A Magyar Rorschach Standard táblázatai*. Schoolbook Publisher. Budapest.

VARGHA, A. – DELANEY, H. D. [1998]: The Kruskal-Wallis Test and Stochastic Homogeneity. *Journal of Educational and Behavioral Statistics*. Vol. 23. No. 2. pp. 170–192.

VARGHA, A. – DELANEY, H. D. [2000]: A Critique and Improvement of the CL Common Language Effect Size Statistic of McGraw and Wong. *Journal of Educational and Behavioral Statistics*. Vol. 25. No. 2. pp. 101–132.

WILLIAMS, B. – MANDREKAR, J. N. – MANDREKAR, S. J. – CHA, S. S. – FURTH, A. F. [2006]: *Finding Optimal Cutpoints for Continuous Covariates with Binary and Time-to-Event Outcomes*. Technical Report Series No. 79. Department of Health Sciences Research, Mayo Clinic. Rochester.