Telmar Centab® - the new analytical possibilities with coefficients of structural associations

Igor Mandel, Telmar Group Inc.,

imandel@telmar.com

Abstract

Telmar Group Inc. has created and made available for the first time a single respondent database Telmar-Centab based on the most reliable source of demographic information – annual Public-Use Microdata Samples (PUMS) data by the US Census Bureau. It has 3 million of respondents, distributed across 3,000 counties and 150 demographic variables, covering all traditional metrics used in advertising media planning. It could be merged on a fly (by Telmar's "multibasing" technique) with other major sources of marketing information. Telmar-Centab provides the explorer with unique opportunity to cross-tab a huge number of variables and make further analysis. It also opens the opportunities for implementation of the new promising analytical approaches (in particular, the recently proposed coefficients of structural association, CSA). It is shown, that CSA bear important features, not observed in traditional coefficients of association, and in particular situations - for binary variables - coincide with phi-coefficients (which are equal to Loevinger's coefficients in this case).
Kew words: advertising; media planning; demographic database; Telmar; Census; Loevinger and phi coefficients of correlation

1. Introductions

In (Mandel 2011) was shown how many new analytical possibilities are open in Centab with three millions of respondents. Also, there were proposed a special measure of association between qualitative and quantitative variables, based on estimation of the differences between shares of values exceeding some typical level (like median) between groups determined by qualitative variable. These ideas were elaborated much further in (Lipovetsky, Mandel 2012), referred below as LM12, where general concept of the coefficients of structural association was developed. This article overviews results from LM12; all proofs could be found there.

Various measures of correlation and association between statistical variables had been introduced about a century ago by F. Galton, K. Pearson, C. Spearman, and over the time many other authors have brought their input into the development of numerous other measures and their interpretation (see, for instance, E. Pearson, 1928; Sheskin, 1997; Rodgers and Nicewander, 1988; Mirkin, 2001; Warrens, 2008). Such measures comprise an essential part of any statistical data analysis and modeling. The current work considers a new measure of a y variable dependence upon another variable x, and this measure is based on the distribution of y along the segments formed by x. These segments are defined due to the structure of the data in x, for instance, if it is a nominal variable the segments naturally correspond to its categories. The profiling of y across x yields an index of association similar to the coefficient of determination, or correlation ratio, well known in the linear and nonlinear regressions and in the analysis of variance. Constructing such a coefficient of association also requires a setup of an appropriate target which can be reached in the best case. For instance, in regression modeling, such a target is presented by a theoretical model which yields the maximum of the estimated

theoretical variance (minimum of the residual variance) so the coefficient of multiple determination will be close as possible to one. In this problem the target can correspond to the maximum possible variance, and quotient of the empirical and the target variances yields a characteristic which called in LM12 the Coefficient of Structural Association (*CSA*). The term structural association reflects that *CSA* expresses internal correspondence between the variables via accounting for the structure of the dependent variable distribution by the segments of the independent one.

2. The Measures of Structural Association

For explicitness of the explanation, at first let us present the suggested approach on a simple example of measuring income dependence on the level of education. Table 1 presents some descriptive statistics based on a sample of 189,771 respondents taken from the data on 2009 Census for people of the age 25 or older, with income above $1,000 per year, and having some level of education. The data is taken from the American Community Survey (see detail in Mandel, 2011). The median income equals $30,000. Table 1 shows proportion of the respondents by the four levels of education, the mean income in each level, and percent of those with income equal or above the median income within each level of education (the total of the last proportion is slightly above 50% because of the cases with values equal the median).

Table 1. Income by level of education

Characteristic	Graduated College and Higher	Attended College	Graduated High School	Did Not Graduate High School	Total
Respondents, %	11.4	47.1	28.5	13.0	100.0
Average income, $	87,721	47,970	29,445	20,073	43,611
Earning more or equal to the median income ($30K), %	82.6	60.1	37.6	18.3	50.8

Considering how strong a person's education determines the level of income, one can apply a common measure of the so-called correlation ratio η^2 (Eta-squared) estimated as a quotient of the weighted between groups variance to the total variance. Similarly to the well-known coefficient of determination R2 for regression, η^2 belongs to [0, 1] interval, is closer to one for a stronger dependence, and reduces to the pair linear correlation squared for the simple case of two groups. For the data in the example above, the Eta-squared estimated by the four groups of education is $\eta^2 = 0.125$. It seems to be a rather low value to justify a general claim that "Studies by the US Census Bureau and many other agencies have consistently shown that people with a higher level of education make more money than those with less education"(Education and Income, www.education-online-search.com/articles/special_topics/education_and_income).

Actually, one can see by Table 1 that there is a substantial difference between the groups, and a qualitative analysis can be straightforward as follows: nationwide, a half of all people has annual income less and a half has more than $30,000; if you had not completed high school you would fall into the group where only 18% have income more than that level; but if you had got a college degree or higher education you would belong to the group where 83% of people have income higher than the median level. The association between the two variables is visible and strong (and reminding roughly 80/20 Pareto rule- see (Lipovetsky, 2009)). Thus, the low value of η^2 does not seem to be an adequate measure for expressing the relation between income and education levels, and in such situation suggested in LM12 the simple and reliable measure of a **coefficient of structural association, CSA** could be a good alternative (I will return to this example later in the Section 3).

Suppose there is a dependent variable *y* structured by *K* groups of the independent variable *x*. Taking *y* at the mean level, or median, or another percentile important as a target value, one can find the frequency value p_i of *x* reached within each i-th group *(i=1,2,...,K)*. For instance, such frequencies are presented in the last row of Table 1 for *K*=4 where the target value is the median. Than one can calculate the average value and sample unbiased variance *Vsample* of these empirical frequencies. For *K* frequencies one may consider a theoretical maximum possible variance which can be reached if these frequencies could have any values on the [0, 1] interval. Having the maximum possible value for the unbiased sample variance Vmax one can estimate a new convenient measure of the dependence of *y* by *x* as a quotient of the variance *Vsample* to the maximum variance *Vmax*. Let us define this CSA via the relative maximum variance:

$$\xi_M^2 = \frac{V_{sample}}{V_{\max}}, \quad 0 \le \xi_M^2 \le 1 \tag{1}$$

If the frequencies *pi* are close one to another, then *Vsample* is small, and CSA is close to zero. If the frequencies vary by groups and are maximally different, then of course CSA reaches one. It is proved in LM12 that the maximum variance corresponds to the frequencies tending to the margins of the interval [0, 1], with about a half of them at each border. The solution can be presented as:

$$V_{\max}(K) = \begin{cases} \frac{1}{4}\left(1 + \frac{1}{K}\right), & for\ K\ odd, \\ \frac{1}{4}\left(1 + \frac{1}{K-1}\right), & for\ K\ even. \end{cases} \tag{2}$$

It is clear that for large *K* the variance (2) has the asymptote value 0.25, and the maximum standard deviation reaches 0.5. Table 2 presents for several *K* the unbiased estimate for the maximum variance and the standard deviation (STD). It is interesting to note that the variances for an odd number of points and the next even number of the points are the same. By the pattern of a sequence for the variances given as quotients it is evident how the table can be continued.

Table 2. Maximum variance for independent frequencies by the numbers of groups.

K	2	3	4	5	6	7	8	9	10
V_{max}	1/2	2/6	2/6	3/10	3/10	4/14	4/14	5/18	5/18
STD_{max}	0.707	0.577	0.577	0.548	0.548	0.535	0.535	0.527	0.527

The closed-form solution for two categories and *q*=0.5 (the median) proportion of the total sample can be expressed via the free parameter of the first weight γ_1 :

$$V_{\max}(\gamma_1) = \begin{cases} \frac{1}{4}\left(\frac{1}{1-\gamma_1}-1\right), & for\ \gamma_1 \le 0.5, \\ \frac{1}{4}\left(\frac{1}{\gamma_1}-1\right), & for\ \gamma_1 > 0.5 \end{cases} \tag{3}$$

The maximum possible variance across all γ_1 can be reached for $\gamma_1 = .5$, when it equals V=0.25 that coincides for the asymptote to the result (2). Table 3 the estimates (3) for the maximum variance and the standard deviation.

Table 3. Maximum variance for two groups and different values of the weight γ_1.

γ_1	0	.100	.200	.300	.400	.500	.600	.700	.800	.900	1
Vmax	0	.027	.062	.107	.166	.250	.166	.107	.062	.027	0
STDma	0	.166	.250	.327	.408	.500	.408	.327	.250	.166	0

A general case of K weighted connected frequencies p_i (i.e. when groups sizes are not equal to each other and respectively the maximal variance should be calculated based on that fact) is defined in (4):

$$p_1\gamma_1 + p_2\gamma_2 + ... + p_K\gamma_K = q \tag{4}$$

where γ_i are the weights of categories (groups), and *q* is equal to the total frequency targeted (for instance, q=0.5 for the median). The solution for the maximum possible variance can be obtained by means of the quadratic programming with linear restriction. Due to the Kuhn-Tucker theorem of the nonlinear programming, as shown in LM12, the global maximum for the variance can be reached for the frequencies *pi* tending to the margins of [0, 1] domain of their possible values.

For a simple case of two connected frequencies the point of the maximum variance is reached for *q=(p1+p2)/2*, so for the equal weights. For equal weights ($\gamma_1 = 0.5$) the maximum variance is defined by the maximum of two values, q^2 and $(1-q)^2$, which presents a maximum squared distance of p_1 and p_2 from their center, q. If the value q is close to the median, so approximately equals 0.5, then both side distances q^2 and $(1-q)^2$ are similar, and the two frequencies are located in the vicinity of zero and one, and the maximum variance value corresponds to the points at the margins of their domain. Even being restricted by (4) the frequencies pi in the point of optimum are far from their center and close as possible to 0 or 1. Thus, for similar weights γ_i and *q* about 50%, the maximum variance obtained in the point of optimum pi by the nonlinear programming can be roughly approximated by the closed-form solution (2). It is important to indicate that in a case of the variables of a special nature the CSA can be related to some other measures of the dependence between them. For instance, in the case of two binary variables *x* and *y*, the so-called Loevinger's H-coefficient presents a better measure of their correlation than Yule's association coefficient or Pearson's phi-coefficient of correlation because it also takes into account the range of the possible marginal values of

frequencies (see Loevinger, 1948; Warrens, 2008). The Loevinger coefficient can be obtained after normalization of the phi-coefficient by its maximum value. The relation between these coefficients can be written as:

$$r_{adj}^2 = H_{Loevinger} = \xi_M^2 . \quad (5)$$

It shows that in a simple case of two binary variables CSA actually coincides with the classical measure of the variables relation.

In another important case of the numerical *y* and the ordinal *x* variables, the CSA characteristic can be defined by a more adequate measure. For an ordinal *x* variable, the reached frequencies by categories are concentrated not at the borders, but rather equally distantly on the interval [0, 1] (for instance, see the last row in Table 1). See details of this scenario in LM12.

Concerning a possibility of the statistical comparison of a sample variance with the maximum variance, there could be suggested the following approach. Let us take a characteristic of the inverted CSA (1) which is defined as the larger (maximum) variance divided by a smaller sample variance, so it is the F-statistics with both degrees of freedom equal *K-1* in unbiased estimation of both variances:

$$F_{K-1,K-1} = \frac{V_{\max}}{V_{sample}} = 1/\xi_M^2 . \quad (6)$$

Similarly F-statistics can be defined for the ξ_E^2 index (6) as its reciprocal value. Table 5 presents examples of such F-statistics for several degrees of freedom *K-1* and levels of significance α. If a calculated F-value (6) is larger than the corresponding critical value from Table 5, then with the confidence probability $1-\alpha$ the sample variance is significantly less than the maximum possible variance, so in this case the sample distribution of the frequencies by categories differs much from the pattern corresponding to the maximum variance.

Table 5. Critical values of FK-1,K-1 statistics for several degrees of freedom K-1 and levels of significance α

α \ K-1	1	2	3	4	5	6	7	8	9
0.1	39.9	9.0	5.39	4.11	3.45	3.05	2.78	2.59	2.44
0.05	161.4	19.0	9.28	6.39	5.05	4.28	3.79	3.44	3.18
0.025	647.8	39.0	15.44	9.6	7.15	5.82	4.99	4.43	4.03

Thus, the dependence of *y* on *x* in this case is not substantial. And vice versa, if the sample F-value is smaller than the one given in Table 5, then one cannot distinguish the sample and maximum variance, so the relation between *y* and *x* can be considered as a strong dependence.

3. Numerical examples

Let us return to the example presented in Table 1. The unbiased estimation of the sample variance by the four frequencies by categories shown in the last row equals 7.74%, and the maximum variance from Table 2 for *K=4* equals 33.3%, so ξ_M^2 (1) equals 23.2%. Considering the maximum variance with the weights of the groups (those

are in the first row of Table 1) and solving the nonlinear programming problem for it I obtain the value Vmax=18.4% (reached at the point $p_1=0.02$, $p_2=0.16$, $p_3=0.99$, $p_4=1.0$, so with values close to the margins of the 0-1 interval), so ξ^2_M equals 42.1%., almost twice than in the estimation without accounting for the groups' weights.

Another example using a subsample from the data on income and education is considered in Table 6. It presents data on income and its cumulative frequency, together with frequencies by the categories of education. Each level of income cumulative frequency grows faster for the lower than for a higher education. In each row of Table 6, by the attained values of the four frequencies by categories the unbiased variance is calculated. Due to Table 2, the maximum variance for $K=4$ equals 2/6, so the relative variance (1) shown in Table 6 is three times higher than the sample variance, and the corresponding values of the sample F-statistics are given too. The last two columns in Table 6 present the index (7) and the corresponding F-values. Incomes up to and above 60 \$K define the median of the total distribution. At the same time this level comprises about 80% and 20% for the "no school" and "college" education, respectively.

Therefore, a half of population earns up to \$60K, and a four fifth of uneducated population belong to this half. And another half of population earns above \$60K, and a four fifths of educated population belong to this other half. At the level of \$60K income, the maximum ξ^2_M = 21.1%, and the statistics (8) hits its minimum F=4.75. Also at this level, ξ^2_E = 37.9%, and F=2.64. Due to Table 5 for K-1=3 degrees of freedom, the critical value for $\alpha = 0.1$ is F=5.39, and it is even more for a smaller alpha. Thus, the sample F-statistics are smaller than the critical one, and one can conclude that the sample and maximum variances in the frequency distributions across the categories of education are undistinguishable, so the income is strongly related to the level of education in the vicinity of the median income frequency. Scanning by the rows of Table 6 shows that this relation diminishes for the lower and higher values of income.

Behavior of these characteristics is illustrated in Figure 1 which shows that maximum values of indices are reached around the target value, the median.

Figure 1. Distribution of coefficients of the relative structural dependence

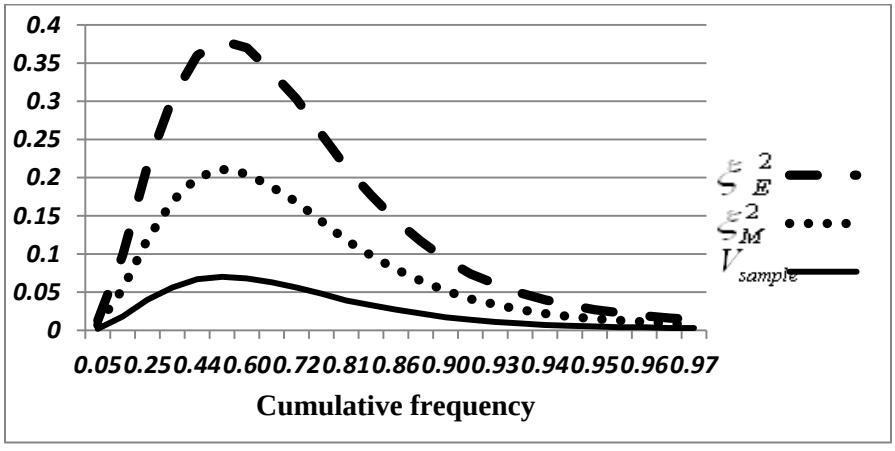


Table 6. Cumulative distributions by education categories, and their statistical characteristics

Income		Cumulative frequency by education category					Relative to maximum variance	
$K	Cumulative frequency	college	some college	school	no school	Unbiased variance	CSA	F-statistic
10	0.050	0.012	0.036	0.059	0.125	0.002	0.007	142.42
20	0.143	0.033	0.101	0.179	0.346	0.018	0.054	18.35
30	0.246	0.061	0.182	0.322	0.525	0.040	0.119	8.38
40	0.345	0.101	0.273	0.451	0.653	0.056	0.169	5.93
50	0.438	0.151	0.364	0.566	0.751	0.067	0.200	5.00
60	0.523	0.210	0.451	0.665	0.820	0.070	0.211	4.75
70	0.596	0.273	0.531	0.742	0.872	0.068	0.205	4.87
80	0.662	0.337	0.607	0.806	0.904	0.063	0.188	5.32
90	0.718	0.397	0.673	0.854	0.929	0.056	0.168	5.95
100	0.766	0.462	0.730	0.891	0.949	0.048	0.143	7.01
110	0.805	0.521	0.778	0.918	0.961	0.039	0.118	8.47
120	0.837	0.572	0.816	0.936	0.972	0.033	0.098	10.20
130	0.863	0.619	0.848	0.950	0.978	0.027	0.080	12.52
140	0.883	0.659	0.873	0.960	0.982	0.022	0.065	15.33
150	0.900	0.699	0.893	0.968	0.985	0.017	0.052	19.34
160	0.914	0.733	0.909	0.974	0.988	0.014	0.041	24.31
170	0.926	0.760	0.923	0.978	0.990	0.011	0.034	29.64
180	0.935	0.785	0.933	0.981	0.991	0.009	0.027	36.75
190	0.942	0.807	0.942	0.984	0.992	0.007	0.022	45.48
200	0.949	0.825	0.949	0.986	0.993	0.006	0.018	55.19
210	0.954	0.841	0.955	0.988	0.994	0.005	0.015	66.30
220	0.959	0.855	0.96	0.989	0.994	0.004	0.013	79.40
230	0.963	0.867	0.964	0.990	0.995	0.004	0.011	94.08
240	0.966	0.877	0.967	0.991	0.995	0.003	0.009	109.28
250	0.969	0.885	0.970	0.992	0.995	0.003	0.008	125.29

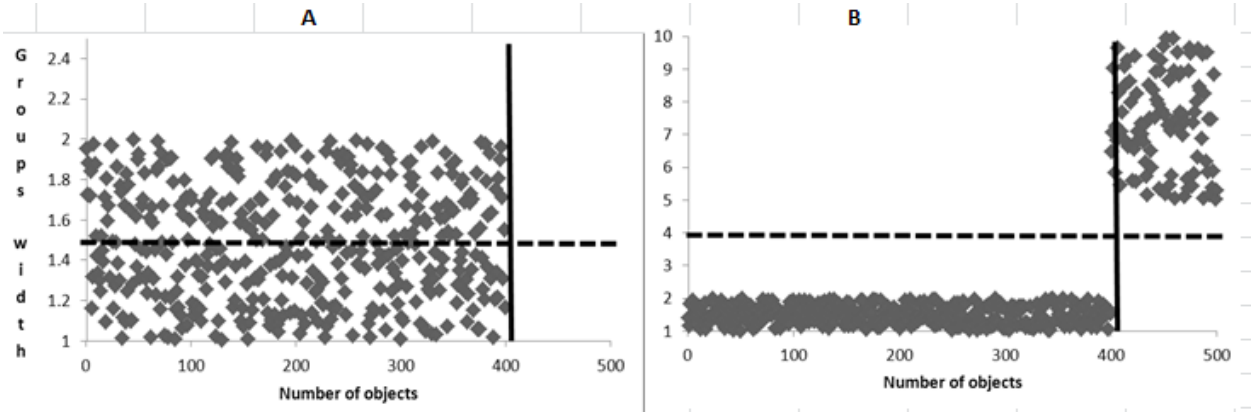
4. CSA Simulation Results for Nominal-Numerical and Two Numerical Variables

The case when x and y are the nominal and numerical variables, respectively, is the most appealing for motivating CSA. Let us start with x as a binary variable when the correlation ratio Eta-squared η^2 equals the squared coefficient of linear correlation. The maximum weighted variance between two groups is defined via the weight of the first group and a targeted frequency q of the whole sample. In LM12 we have conducted the simulation experiment using generated datasets to study CSA calculated via the

maximum variance (3) in comparison with η^2 . Data simulation and experimental design are as follows. The data for numerical y was generated as uniform random values for two groups corresponding to 1 and 0 values of the binary x, respectively. The proportion of y-values in the 1st group for x (those equal 1) was taken as 20%, 40% and 90% of the total sample size of 500 values. Values of *y* in the 1st group are distributed in the interval [1, 2]. Values of *y* in the 2nd group for *x* (those equal 0) are distributed in the interval [a, b] with different margins a and b, and different width of b-a =1, 3, and 5 (which also corresponds to different variances (a-b)2/12). The *y*-values in the 1st and 2nd are either overlapped or distanced. The overlapping was taken at 70%, 50%, or 0% of the width of the 1st group (which equals one). A distance between groups (measured as a gap between the right margin of the 1st interval and the left margin of the 2nd interval) was taken equal to one or three widths of the 1st group.

For example, Figure 2A presents a scatterplot of the dataset with the 20% proportion of y-values in the 1st group, overlapping 50%, and the width of the 2nd group equals 1 (so a=1.5 and b=2.5). Figure 2B shows the dataset with the 90% proportion of y-values in the 1st group, with width a-b=5, and the gap of 3 units (so a=5 and b=10).

Figure 2. A: Overlapping groups of the same widths; B: gapped groups of different widths.



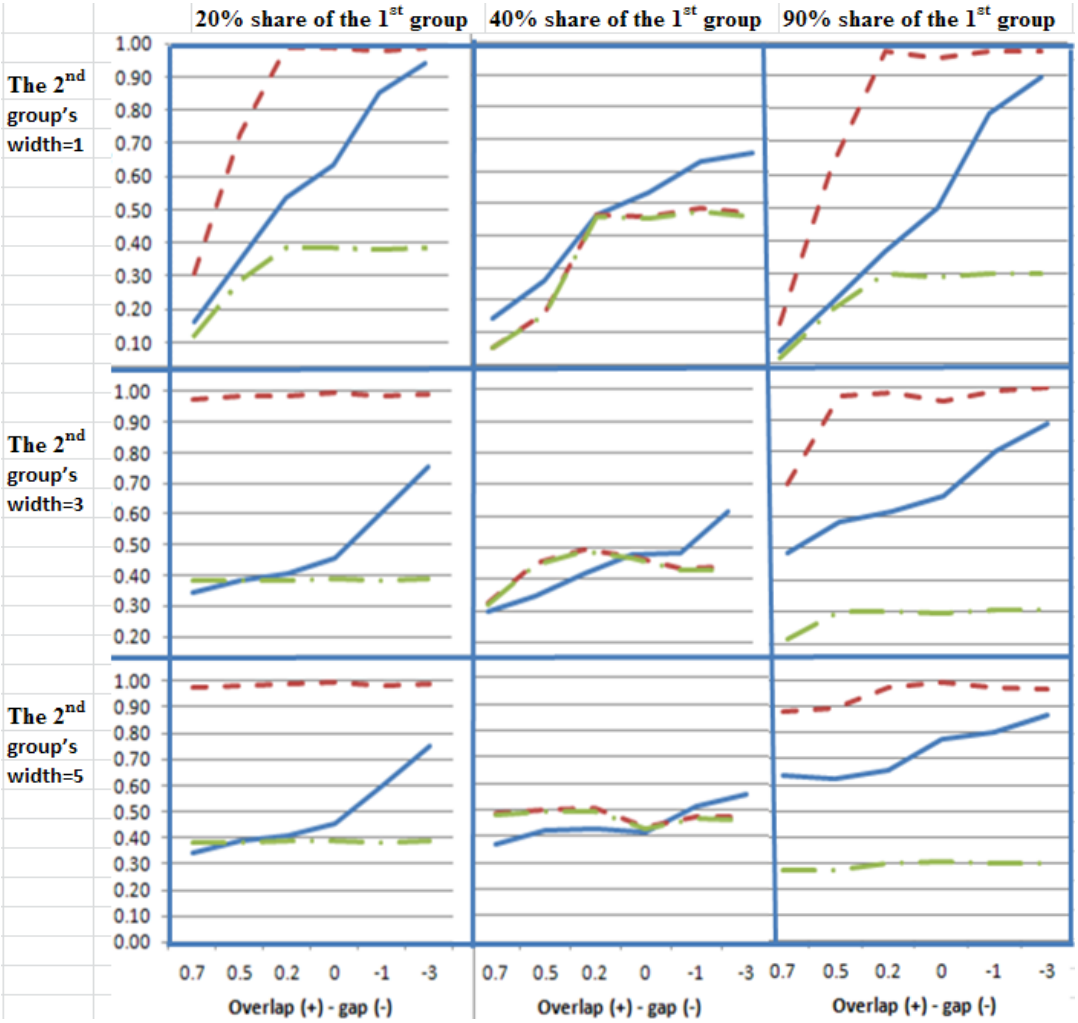
For each of the mentioned combinations of experimental parameters were compared the three measures of association: η^2 , CSA (2), and also the un-weighted CSA (discussed below). Due to (2), the maximum variance for the case of two groups, K=2, equals 0.5. The results of the experiments are summarized in Figure 3 where each panel consists of estimations by six datasets. For instance, Figure 2A dataset's results are presented at the second point in the abscissa in the left upper panel, and Figure 2B dataset's results are reflected at the last 6th abscissa point in the right lowest panel of Figure 3. Actually, with ten random drawings for each of the six types of datasets, each panel in Figure 3 presents results evaluated by sixty numerical simulations. Several observations can be made by the experiments.

1. Eta squared η^2 increases monotonically, but the slope and interval of its change varies. For instance, η^2 for the data in the upper left panel and in the lower right panel of Figure 3 are about 0.25 and 0.87, respectively.

2. Looking at the graphs in Figure 3 from the upper left to the lower right corner η^2 ended up at about the same value 0.9 though starting from very different initial values from about 0.1 to about 0.6. Thus, for η^2 comparison the relative size of the groups should be taken into account. For two partitions of the similar relative sizes (20-80% like

in Figure 2A, and 90-10% like in Figure 2B) with the same gap between groups (the lowest row in Figure 3) the value of η^2 could be as different as 0.3 versus 0.6.

Figure 3. Comparison of Eta-squared and CSA for different data settings.



Eta squared – solid line; CSA – dashed line; CSA un-weighted – dashed-dotted line.

3. CSA typically shows values higher than η^2 for any datasets; they become both close to zero when groups have the same sizes and variances, i.e. CSA is more sensitive to deviations from symmetry in the distribution of y-variable within groups.

4. When variances in groups are similar, the CSA reacts to the level of overlapping more steeply than η^2 regardless of the relative sizes of the groups (see the curves in the 1st row of the panels in Figure 3, and to a lesser degree in the 2nd row there). But in contrast to η^2, CSA reaches the maximum value when groups are not overlapped at all, while η^2 grows further if the gap between the groups widens.

5. When variance within the group with larger y-values is noticeably greater than variance for the group with smaller y-values (those of x=0 and x=1, respectively) CSA is

almost always very high (rows 2 and 3 in Figure 3), with a slight dependence on the degree of overlapping. It holds until the share of the 1st group becomes larger than 50% when a slight decrease in CSA can be seen with overlapping (the curves in the lower right corner in Figure 3).

6. CSA is defined by the proportions and weights, not by the actual numerical values, and for this reason it is a robust statistics. While r^2 may be severely affected by outliers, the CSA remains intact that makes it a good tool for data analysis.

7. The un-weighted CSA is useful in situations when the relative sizes of the groups are not available, and only proportions of the exceeding target's values in each group p_i are given. Behaving similarly to the regular CSA, the un-weighted CSA is defined by the unbiased variance of the proportions and the maximal variance (2). The correlations between these two estimates of CSA are very high (see any panel in Figure 3) that suggests a possibility to estimate the CSA without the original data, only by usually available proportions, but it needs additional investigation.

For a general case of many groups, K>2, it is possible to use evaluation without weights (2), but there is no analytical closed-form solution for the maximum variance estimated by weighted proportions. The latter can be obtained by means of quadratic programming or by maximization algorithms available in modern mathematical and statistical software packages.

Let us consider now the case of two numerical variables *x* and *y*, with *x* sorted in the ascending order. The entire interval [xmin , xmax] is divided into *K* roughly equal groups (so each group has about N/K values where N is a sample size); the target statistics for y within each of these groups is calculated together with CSA estimation. It is clear that if the variables are highly correlated the CSA would be high as well. For example, if one considers the median of y as the target, about a half of the groups should have zero frequencies p_i (because almost all values in these groups are less than median after sorting by x), and about half would have frequencies of the value one. Then the variance of frequencies should be close to its theoretical maximum, CSA is close to 1, and in this case one can use the formula for maximum variance (2).

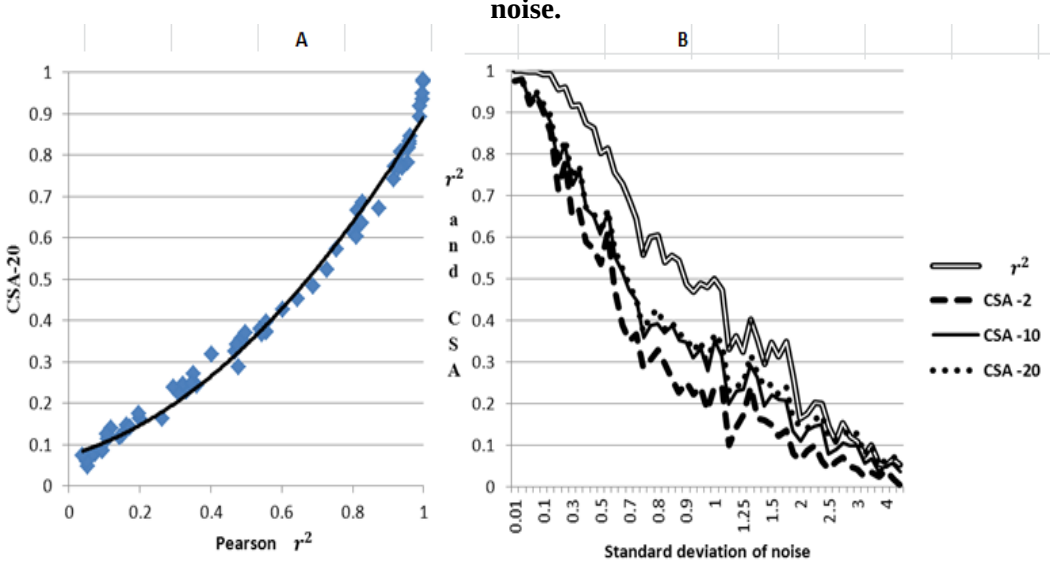
To look closer at this estimation, it were designed the following experiment (LM12). Data by *x* was generated as a normally distributed variable (N=4,000) and sorted. Data by *y* equals *x* plus a random normal noise of zero mean and different levels of variance. Then the data was divided into different number of groups: 2, 4, 5, 10, and 20 groups, and the CSA were calculated with (2) for the maximum variance (let's denote them CSA-2,…, CSA-20, by to the number of groups). For each level of noise, ten random draws were taken, and the average CSA, Pearson linear, and Spearmen rank correlations were calculated. The noise variance varied so that the linear correlation between variables changed from 0 to 1, which allowed us to observe the related CSA values. To evaluate the CSA robustness data were distorted y by various kind of outliers and their clusters to create different situations which show comparative behavior of the three considered measures of association. The experiment demonstrated that CSA reveals certain features in the data which cannot be detected by the traditional measures.

Let us consider results of comparison between CSA and Pearson's correlation (Spearman's rank correlation behaves very closely to Pearson's, so is not shown). Figure 4, A and B, presents characteristics profiled by the noise's standard deviation σ, and shows the following.

1. In general, CSA behaves similarly to the linear correlation squared. The correspondence between the Pearson's r^2 coefficient and CSA-20 is shown in Figure 4A on the scatter-plot. They are closely related, with the coefficient of determination equals

0.99 for the quadratic model, so CSA provides with essentially the same information on the strength of the relationship as r^2 does.
2. Linear correlation squared is usually higher than CSA. For CSA-20 a typical difference lays within 0.1-0.2 (see Figure 4B).
3. CSA value grows with increase of the number of groups – the difference between the CSA-2 and CSA-20, with 2 and 10 groups, respectively, is especially noticeable.

Figure 4. Correlation and CSA: A. CSA-20 versus r2; B. r^2 and CSA profiled by noise.



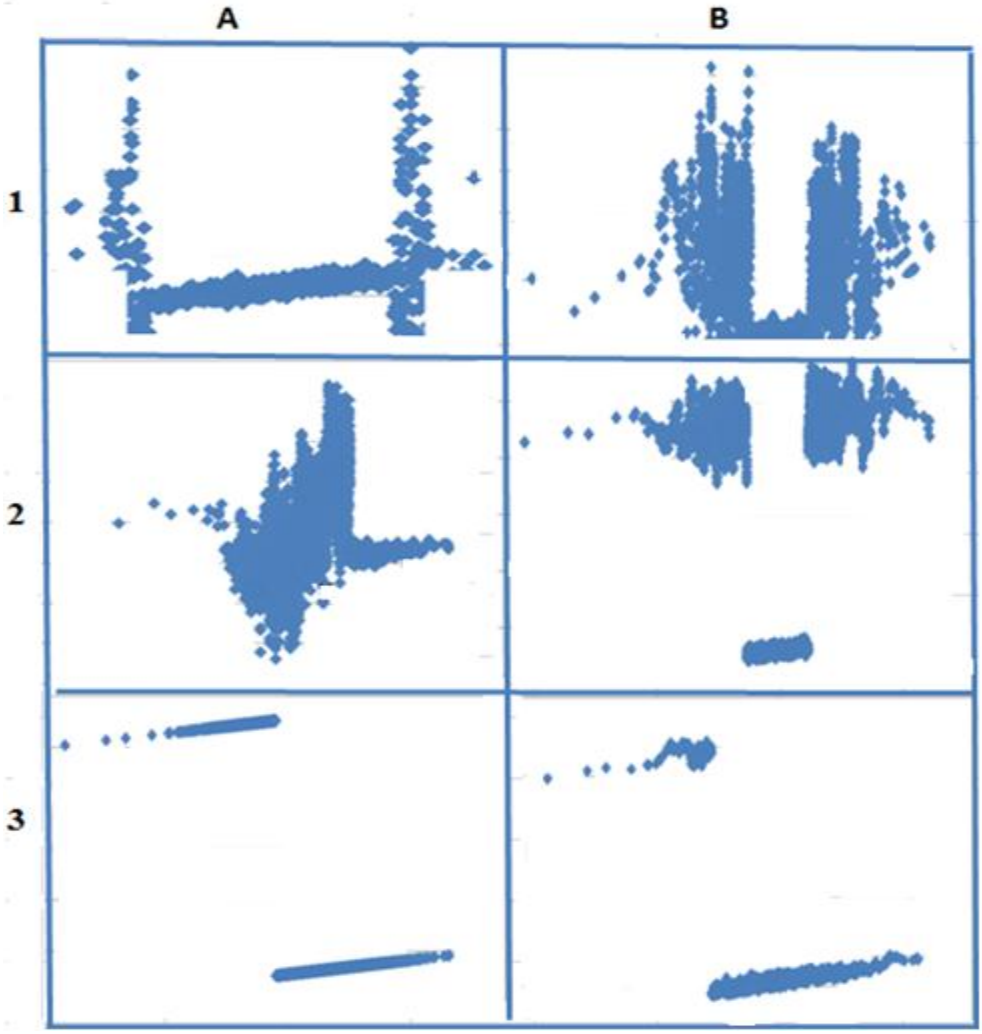
The most interesting observations obtained by the variety of experiments with different datasets are summarized in Table 7 which is also illustrated by Figure 5 with some specific data patterns. Each of these patterns shows that two or three of the used coefficients disagree.

By Table 7 and Figure 5 one can conclude the following.

1. The CSA rather corresponds to the rank than to the linear correlation. If the rank correlation is noticeably different from the linear correlation, the CSA would also differ from the linear correlation. For instance, the rank correlation is high but the linear correlation is low (the 1st row of Table 7), which indicates that a strong non-linearity or a large impact of the outliers reduce the linear but retain the rank correlation. Another common situation is when a single outlier (or a small group of outliers) makes a high linear correlation to become low, while the rank correlation and CSA remain high (the 2nd in Table 7). The CSA and rank correlation are powerful indicators of the deviation from the linear relationship.

2. There are situations when CSA outperforms the rank correlation. Two such examples are given in Table 7, row 3 (B1 and A2 in Figure 5). They occur when both Spearman and Pearson coefficients are low but CSA is high in the presence of some large groups of intermixed data: there is a strong non-monotonicity when the rank correlation does not work in B1; in A2 is a mixture of clusters. Another case is shown in B2 with three large clusters: data in the middle one has a strong positive correlation, but in two others do not. Yet their remote locations and sizes (25% of the sample in each one) guarantee that the share of values exceeding median varies substantially in the direction of *x* values, resulting in a high CSA.

Figure 5. Data patterns detected differently by CSA, linear and rank correlations.



Columns (A, B) and rows (1, 2, 3) are named for a convenient referencing.

3. Another type of the relation is shown in row 4 of Table 7 (A3 in Figure 5). A small rank correlation occurs together with a high negative linear correlation which is explained by the mutual locations of two clusters that produces a completely misleading impression about the data structure. But a high CSA shows that the data is ideally aligned in a sense that a vertical shift between the clusters does not affect the distribution of *y*-values exceeding the median while moving along the *x*-axis. This example is especially interesting because both the rank and linear correlations give a distorted view of the data, although by completely different reasons (due to the shift in values and due to the mutual location), while the CSA captures the important fact that the variables are strongly (yet piecewise) correlated. It may happen in situations when a data trend changes abruptly (like after a financial crisis), and then continues in the same fashion as before but starting from a lower point.

Table 7. CSA, Pearson and Spearman coefficients as indicators of anomalies in data.

	Data structure (reference to panels in Fig. 4)	Numerical values			Correlation pattern		
		Rank	Linear	CSA	Rank	Linear	CSA
1	Two groups of 5% outliers, positive dominant correlation (A1)	0.61	0.07	0.70	+	**0**	+
2	5% of outliers with highly correlated data (B3)	0.60	-0.38	0.69	+	-	+
3	Highly non-monotone relation (B1), messy data with large intermixed correlated cluster (A2), or two uncorrelated large clusters with correlated main dataset (B2)	0.16	0.02	0.89	**0**	**0**	+
4	Two positive within-group (25%-75% shares) correlations vs. negative total correlation (A3)	-0.13	-0.72	0.89	**0**	-	+

5. Summary

There are considered new types of the measures for estimation of the variable dependence on another one. The suggested Coefficients of Structural Association, CSA, are based on the distribution of one variable across the segments formed by another one. The quotient of the sample variance of the frequencies across the segments by the maximum possible variance constitutes the index of relative maximum variance. This index has features similar to the coefficient of determination in regression modeling, or in the analysis of variance. Finding of the maximum possible variance is considered in the cases of independent and dependent shares of the segments.

Coefficients of Structural Associations have powerful interpretational capability, and serve as viable indicators in data analysis for detecting the abnormalities which cannot be found by the traditional measures of correlation. They also demonstrate a close relationship with other measures of association, showing in certain cases equivalence with them (the Loevinger's coefficient for binary variables).

The considered methods enrich both theoretical and practical estimations for identifying variables relations hidden in the data. Future studies are needed, both theoretical and experimental, to clarify the CSA abilities in data analysis with different scales, and for multivariate data and its modeling.

References

Lipovetsky S. (2009) Pareto 80/20 Law: Derivation via Random Partitioning, *International J. of Mathematical Education in Science and* Technology, 40, 271-277.

Lipovetsky S. and Mandel I. (2012) Coefficients of Structural Associations International *Journal of Information Technology & Decision Making, in progress*

Loevinger J.A. (1948) The technique of homogeneous tests compared with some aspects of scale analysis and factor analysis, *Psychological Bulletin, 45, 507–530.*

Mandel I. (2011) Demography, Geography and Marketing: Telmar Centab – the largest US Census based study, *Proceedings of the Joint Statistical Meeting, The American Statistical Association, Miami, FL, 2691-2698.*

Mirkin B. (2001) Eleven Ways to Look at the Chi-Squared Coefficient for Contingency Tables, *The American Statistician, 55, 111-120.*

Pearson E.S. (1926) Review of Statistical Methods for Research Workers (R. A. Fisher), *Science Progress, 20, 733-734.*

Sheskin D.J. (1997) Handbook of Parametric and Nonparametric Statistical Procedures. *CRC Press, Boca Raton, New York.*

Warrens M. (2008) On association coefficients for 2×2 tables and properties that do not depend on the marginal distributions, *Psychometrika, 73, 777–789.*