***Cumulative curves***

SPSS macros by Kirill Orlov

kior@akado.ru, ttnphns@gmail.com

<https://www.spsstools.net/en/KO-spssmacros>

All rights reserved

*Cumulative curves*. Macros that are related to analysis of cumulative distributions. One of them comparing, via cluster analysis, subsamples by shape of cumulative distribution in variables. Another macro – for marketing – analyzes data of the so called price sensitivity meter (PSM).

*Read “*[*About SPSS macros*](https://www.spsstools.net/en/KO-aboutmacros)*” what are they and how to run them.*

*The “Protected directory” error.* Some of the macros described in the current document write temporary files to hard disc. If you don't have full Administrator rights of your computer, it may cause error saying, among things: *“SPSS Statistics cannot access a file... specifies a protected directory...”*, meaning that the default directory the macro wants to use is protected on your PC. To solve the problem, in Syntax window issue command: CD 'myfolder'., where 'myfolder' is the path/name of some folder where you are allowed to save files to.

# MACRO !KO\_PRICECRVS: PRICE SENSITIVITY CURVES

Version 3, Aug 2015 (Version 1, Jun 2004). Tested on SPSS Statistics 20, 22, 26.

*This macro needs SPSS Statistics with Custom Tables option*

!KO\_pricecrvs price1= *q14a* S /\*Variable "too expensive price", and after name:

/\*how it was worded - strict (S) or unstrict (M)

/price2= *q14b* M /\*Variable "expensive price", and after name: analogously

/price3= *q14c* M /\*Variable "cheap price", and after name: analogously

/price4= *q14d* S /\*Variable "too cheap price", and after name: analogously

/wbuyer= *0 1 1 1 0* /\*Weights defining categories of buyers: numbers >=0, in number of

/\*number-of-vars+1, in sequence from price4 to price1

/labels= *'Non-buyers "too cheap"'*

*'Buyers "cheap"'*

*'Buyers "normal"'*

*'Buyers "expensive"'*

*'Non-buyers "too expensive"'* /\*Optionally, parallel to WBUYER: labels of buyers /\*categories

/smooth= YES /\*Optional: perform smoothing of the curves - NO (default) or YES

/graph= /\*If not all graphs needed may list specific needed graphs by

/\*letters A, B, C, D, E

/grvar= /\*Optional: make analysis separately for subsamples of this variable

/\*(graphs will be paneled)

/clean= YES /\*Clean input price variables: YES (default) or NO

/nummiss= *1* /\*If to clean: Maximum permitted number of missing responses for a respondent in

/\*price variables (from 0 to number-of-vars -1; default=1)

/numeq= *2* /\*If to clean: How many price variables at maximum may have same response in a

/\*respondent (from 1 to number-of-vars; default=2)

/invers= /\*Where inversion of values found in price variables:

/\*correct it (CORRECT, default) or delete such respondents (DELETE)

/cleansave= *'d:\exercise\cleanvars.sav'* /\*If to clean, optional: save cleaned

/\*variables: path/name of external file

/save= *'d:\exercise\pricetab.sav'* /\*Optional: File to save table of curves, path/name

/ptsave= /\*Optional: File to save prices for the price chiasm points:

/\*path/name.

Minimal specification WBUYER and at least one of PRICE1, PRICE2, PRICE3, PRICE4.

Macro for a pricing research. It builds “pricing curves” which essentially are cumulative distributions of responses to the questions known as *price sensitivity meter*, PSM after van Westendorp. PSM is a popular method, but it suits mostly to goods which are already established on market, the goods for prices in which respondents orientate themselves well.

The four van Westendorp questions which sometimes vary by their wording and order sound approximately like this:

* Above which price will you not buy this product because it would be too expensive? (***Too expensive price***)
* Starting from which price up the product becomes expensive for you, still you would think about buying it? (***Expensive price***)
* Starting from which price down the product would become attractively cheap for you? (***Cheap or Good price***)
* Below which price will you not buy this product because it would be suspiciously cheap? (***Too cheap price***)

Sometimes not all 4 questions are present. The macro is meant for such incomplete format of pricing research too. The macro outputs several price plots. The price curves can be saved by you by specifying the data file to save to.

Before you run the macro, make sure that your input dataset is saved – because the macro spends the input dataset. If you ordered the macro to clean the data, you can save the cleaned input variables as a new file, for subsequent usage (see CLEANSAVE s/c).

EXAMPLE 1.

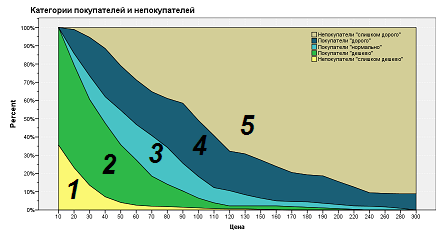
!KO\_pricecrvs price1= q14a S /price2= q14b M /price3= q14c M /price4= q14d S /wbuyer= 0 1 1 1 0

/save= 'd:\exercise\pricetab.sav'.

* Classic PSM study with four price variables; the sample is being layered into 5 categories of respondents – 3 categories of buyers and 2 categories of non-buyers.

**Plot A. Categories of Buyers and Nonbuyers.**

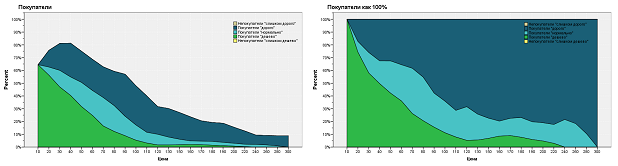
This is the principal plot showing 4 (or whichever the number of price questions) cumulative distribution curves (turned upside down) corresponding – top-down – to the price questions: *Too expensive price, Expensive price, Good price, Too cheap price*. These curves stratify the respondents in four categories by their perception of the price – bottom-up: **1**. *Nonbuyers (“Too cheap”),* **2**. *Buyers (“Good/bargain/for cheap”),* **3**. *Buyers (“Normal”),* **4**. *Buyers (“Expensive”),* **5**. *Nonbuyers (“Too expensive”)*. This graph shows how as price grows the sample structure, its reaction to the price, changes.



On plot A (and also on subsequent B, C, D) some of the cumulates (the curves) are usually shifted by one unit of the price scale to the right, in comparison with the original cumulates. This “correction” of the cumulative curve is introduced depending on how the corresponding price question is formulated – strictly or mildly. The correction for the curves can additionally consist of that if – due to the mentioned shift or due to the deletion in the different price variables of not the same respondents (with bad data) – there occurred some crossing of the cumulates at the edges of the plot (which occurs seldom), – then this decussation is removed by averaging. Finally, with SMOOTH=YES the corrected cumulates are also smoothed.

**Plot B. Buyers. Blot C. Buyers as 100%.**

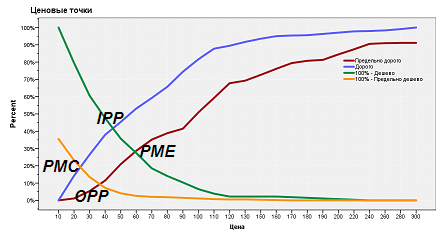
Both plots display only Buyers and are derived from plot A by removing Nonbuyers (layers 1 and 5). Plot C is not drawn if there is only one category of buyers.



**Plot D. Price points.**

If to show two of the cumulates ascending while the other two still show descending there appear four points of chiasm which, by Westendorp’s opinion, set references (this is only one of possible approaches) for decision making about the price for the product or, if that is an old product, say about its status quo on the market.

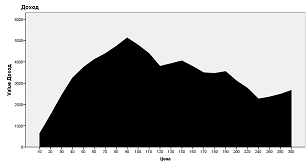
* IPP – point of price indifference.
* OPP – point of price optimality.
* PMC – point of marginal cheapness.
* PME – point of marginal expensiveness.



Plot D is not drawn if the set of pricing questions is such that only ascending or only descending curves could be drawn on the plot. If plot D was drawn the macro prints out price coordinates of the points of chiasm.

**Plot E. Revenue (turnover).**

This plot comes out by multiplying of price level by the amount of buyers at that level (plot B) and it shows the supposed gain in arbitrary unit at different prices for the product.



***Subcommands***

**PRICE1, PRICE2, PRICE3, PRICE4**

The 4 price variables corresponding to the 4 price questions: too expensive (PRICE1), expensive (PRICE2), cheap price (PRICE3), too cheap price (PRICE4). You must indicate at least one of the four. Classic (complete) PSM study presupposes all 4 variables. After each variable name it is necessary to mark by the capital letter how – strictly (S) or mildly (M) – the given question is formulated.

In the example of the questions cited earlier one can see that the 1st and the 4th questions are worded strictly: “above/belowwhich price”; but questions 2nd and 3rd are worded mildly: “starting from which price up/down”. Although in reality the questions may be worded not as pedantic, either strict or mild logic is always assumed, and the macro should be made informed of it. It affects which of the curves it will shift by one price unit to the right. The shift or nonshift by one unit considerably influence the results in case of rough-grained price scale; as it becomes more fine-grained the importance how the question is put – strictly or mildly – attenuates.

*Values and labels*. The variables must be categorical (in the sense - discrete values). If the questions had been open-ended, without graduated price scale of possible responses, then graduate (categorize, bin) the variables prior input them into the macro.

* The values of the input variables may be arbitrary codes or (the discrete) prices themselves.
* If that is *codes* (such as 1, 2, 3, …), then presence of *value labels* in the variables is required – value labels showing prices themselves. The value labels must be the same in all the variables. The labels should be numbers (“200”), not ranges (“100-300”). If these values are fractional (“16.5”), they must be decimal with the separator (dot or comma) meeting the setting of locale of your SPSS Statistics. Currency symbols, such as "$200", "200 р.", are allowed in labels.
* When there are labels, X axis of the graphs will present them all. While when there are no labels (i.e., the values are prices themselves), X axis will depict only those price levels which had responses (nonempty levels).

All irrelevant price levels and other responses (“Other”, “No answer”) should have the status of missing values (user-missing).

**WBUYER**

This subcommand sets which respondent categories (strata on plot A) are buyers and which are not buyers, or, if necessary, weights the categories of buyers differentially. You have to specify the list of nonnegative numbers in quantity 1 more than the number of input price variables, and in the sequence bottom-up – see plot A, – from the side of curve PRICE4 towards curve PRICE1. Zeros indicate nonbuyers, and other values designate buyers. Attention: zero weight, i.e. nonbuyers, write with single digit 0 (not 00 or 0.00).

In the classic price study with 4 questions this list looks as 0 1 1 1 0, i.e. the buyers are the middle three strata of respondents, while two utmost strata are nonbuyers. Sometimes question PRICE4 is formulated in the questionnaire in the sense that it is not the utmost low price for a customer but rather the utmost *just price for a manufacturer*, which does not concern with a customer and presumes that he *will* buy below that price. Then the series of the weights is expected to be: 1 1 1 1 0.

You may indicate different weights for different categories of buyers instead of just ones, for the weighted computation of the Revenue (plot E). For example, if there exists a reason to suppose that “Bargain” buyers will be purchasing 2½ times more often or more pieces than the “normal price” buyers or “expensive price” buyers, one can give them weight 2.5, i.e. 0 2.5 1 1 0. Since Revenue is arbitrary units, only *ratio* of weight values counts.

The revenue (of turnover) curve, plot E, is obtained by multiplication of the respondent category weights with the percent of the given category observed at the given price level, adding up and multiplying by that price level. Percent of the respondent categories are variables *stratum1*, *stratum2*, *stratum3*, *stratum4*, *stratum5* in the dataset which you can save with SAVE subcommand. In the classic PSM study, as already said, there are present all 5 categories and their buying weights are 0 1 1 1 0. Hence, the revenue curve is (0\**stratum1*+1\**stratum2*+1\**stratum3*+1\**stratum4*+0\**stratum*)\**Price*. You may modify this formula if you want to build your own revenue curve using the results file having been saved.

**LABELS**

Here you may indicate the list of quoted labels to name respondent categories on the plots. This list is correspondent to the list WBUYER.

**SMOOTH**

If the plots appear unpresentably jagged you may request smoothing of the corrected cumulates by the moving average with window 3 (SMOOTH=YES).

**GRAPH**

Which plots to produce. By default/unspecification of the s/c the macro returns all 5 plots (A B C D E) if it can. If you don’t want all the plots list here only those you need, for example: GRAPH= A D E.

**GRVAR**

May specify one respondents grouping variable by categories of which the macro will produce the analysis separately. The plots will be of matrix (panel) kind. The groups are the subsamples of respondents (such as age or sex groups). If you have to do the same PSM study for a number of products (versions, competitor brands) but the respondents are the same, - restructure the data from “wide” to “long” format (see Restructure in Data menu) and specify the “product” variable here. However, the macro will treat these as if independent subsamples of respondents: the macro does not do any repeated-measures analysis.

With GRVAR specified the macro will not compute the coordinates of the price chiasm points, even if D plot was drawn.

**SAVE**

You may indicate an external filename to save SAV file containing the table of price curves on the basis of which the plots have been built.

**PTSAVE**

You may indicate an external filename to save SAV file containing the coordinates (prices) of the chiasm points. This subcommand is inactive if D plot is not drawn or when GRVAR is specified.

**CLEAN**

Check/clean input price variables (YES, default), for they could contain data errors/typos, or don’t (NO). See parameters of cleaning in s/c NUMMISS, NUMEQ, INVERS.

**NUMMISS, NUMEQ, INVERS**

These subcommands act if CLEAN=YES (cleaning requested, it is so by default) while price variables are more than one. They set the parameters of cleaning.

NUMMISS is the maximal tolerable number of missing answers in respondent. If exceeded, the respondent is discarded from the analysis. Specify integer from 0 to one less than the number of input price variables. By default NUMMISS=1.

NUMEQ is the maximally tolerable equal responses from a respondent. For instance, among 4 price responses, 80 40 40 20 – two are identical. In responses 80 80 20 20 also two identical (albeit 2 times by two). And in 80 20 20 20 three identical responses. By default NUMEQ=2, that is, a respondent who answered 80 20 20 20 will be discarded from analysis. You may indicate your value for NUMEQ as integer between 1 and the number of input variables.

INVERS checks the entangled data entry. Sometimes it occurs that actually valid respondent’s answers had been somehow mixed up when entered in the price variables, for example PRICE2 contains the response actually pertaining toPRICE1, and the consequence of which is the inversion of the prices: PRICE2 bears price higher than PRICE1, for the respondent. The macro can detect and fix such crosses, moving the values within such respondent so that the inversions disappear. This check/fix is done by default and by INVERS=CORRECT. If you specify INVERS=DELETE respondents with price inversions will be discarded. The macro informs of the amount of inverses discovered.

**CLEANSAVE**

This subcommand acts if CLEAN=YES (cleaning requested) while price variables are more than one. If needed, it will save cleaned price variables on hard disk. Specify path/name of SAV file in quotes or apostrophes.

***Supplements to the macro***

The macro is supplied with graphical templates (several .SGT files). Their usage (optional) will help you save time spent on graphs editing. The templates were optimized for SPSS version 22, they might not fully suit other versions; however, you always could create own templates instead of these by saving a SGT-template from the edited, as you want, graph.

If you want the macro itself to apply the templates then put them in some folder on hard disc, and write down in the macro body in command DEFINE, in argument TEMPLT…!DEFAULT(‘’) the path to the folder. For example: !DEFAULT(‘C:\MY TEMPLATES\’). Don’t forget to end the path by “\”. Don’t change the names of the template files.

### Special regimes

The macro is not meant for split file state of the input dataset, but it obeys case filtering or selection (commands FILTER, SELECT IF, USE). The filtered off cases, although they do not participate in the analysis, still get cleaned (if CLEAN=YES) along with the being analyzes cases. Do not use temporary (under TEMPORARY) operations before the macro run.

*Weighting* the dataset can expand the potential of the price meter research. For example, respondents could be weighted by their propensity to buy the product, which will enhance the predictive quality of the study. Another case: weighting of respondents by the quantity of the product they use to buy or might be expected to buy. Then the estimates of Revenue will become more realistic; only one should bear in mind that under the above mentioned weighting all percents on the plots will reflect not the number of people ready to buy for this or that price but the amount of the product expected to be purchased at this or that price.

EXAMPLE 2.

compute w= npieces\*loyalty.

weight by w.

!KO\_pricecrvs price1= q14a S /price2= q14b S /price3= q14c S /price4= /wbuyer= 1 1 1 0 /clean= NO

/save= 'd:\exercise\pricetab.sav'.

* Besides price variables there was present variable *NPIECES* showing now many pieces of the commodity a respondent purchases per year, and also variable *LOYALY* containing some coefficient of respondent adherence to the brand which the study is about.
* The researcher multiplied these indicators (he might as well have thought out a different function of the two) and weighted the dataset by the computed variable *W*.
* Percents of the cumulative curves will now reflect not the percent of respondents but the percent of respondents weighted by the amount of purchase, i.e. – the percent in terms of product quantity.
* PSM study in this example did not ask PRICE4 question. All the three price questions were worded “strictly”. The buyers are defined as all respondents except the top-most layer (lying above the “too expensive” price). No check/cleaning of the data was performed.

***Related macros***

You can preliminarily reveal groups with surely similar pricing curves by macro !KO\_CUMCLUG (see below) in order to join them in one for the PSM study.

# MACRO !KO\_CUMCLUG: COMPARISON OF GROUPS (BY CLUSTER ANALYSIS) ACCORDING TO THE CUMULATIVE DISTRIDUTION IN VARIABLES

Version 1 (Aug 2004). Tested on SPSS Statistics 13, 14, 26.

!KO\_cumclug vars= *v1 v2 v3 v4* /\*One or more quantitative, with discrete values, variables

/\*with common pool of values between them), name-by-name list

/grvar= *city* /\*Variable setting the groups to compare

/matrxs= CONCAT /\*If multiple variables: average matrices obtained for them in one

/\*(POOL, default), or concatenate in one (CONCAT)

/measure= AUTOCORR /\*Distance to compute between the groups’ cum% curves:

/\*BLOCK (manhattan, default), CHEB (Chebyshev), AUTOCORR (autocorrelative)

/missing= /\*Delete missing values from VARS listwise (LISTWISE) or independently

/\*(VARIABLE, default).

Minimal specification VARS, GRVAR.

The macro takes one or more analyzed variables (non-nominal) and a respondent grouping variable, and returns a square symmetric distance matrix showing differences between the groups in respect to cumulative distribution observed in the analyzed variable(s). The macro also builds a cluster dendrogram (agglomeration method: farthest neighbour) on the basis on the obtained matrix, displaying the said differences. The macro has a sense to apply if there are more than 2 groups.

If analyzed variables are more than one, one can demand to average results (matrices) obtained by them, or to “merge” (append) the results – to get the combined matrix, the single dendrogram depicting differences separately by each variable.

The matrix is output as a new unnamed dataset. Column/row names in it and, correspondingly, labels on the dendrogram contain the group codes: the values of the grouping variable, for example, gr\_10 – the group coded as 10 in the data. You may save and use the matrix in your further analyses.

***Subcommands***

**VARS**

Indicate name-by-name variables which values’ cumulative distributions are to analyze. These should be variables with discrete ascending values, - variables for which a table of frequency distribution does not lacks sense. The values themselves may be any – not necessarily ordinal codes 1, 2, 3, … If variables are multiple, they all must have the same common bank, i.e. scale, of possible values.

**GRVAR**

Indicate numeric categorical variable defining the groups. Each unique valid value in it forms a group of respondents. It is better to use integers (see limitations, below). No negative values are allowed. Cases missing on GRVAR will be filtered out by the macro.

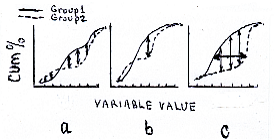
**MEASURE**

Choose the kind distance to compute between the groups:

BLOCK – (default/unspecifying) Manhattan distance; it is the sum of absolute differences between the two being compared cumulative curves (see the picture, a). Use if any discrepancies of the curves, observed along the values scale, are important to you as the difference.

CHEB – Chebyshev distance; it is the maximal absolute difference between the two being compared cumulative curves (picture, b). Use if the curves differ strongly for you, if they are far noncoincident at least at one point.

AUTOCORR – autocorrelative distance; it is the sum of differences between the two being compared cumulative curves, the differences checked by their corresponding lag1-preceding differences (picture, c). Use if only stable differences are important to you – when one curve is above the other at a considerable length.



The autocorrelative distance can, in principle, take on negative values. It happens when one curve crosses the other regularly. With cumulative distribution curves, that is rarity (perhaps unless the scale of possible values is scarce in variables). The macro informs if negative distances have appeared, and abolishes their sign. So, they will come out to the matrix as positive too.

**MATRXS**

This subcommand is in effect if VARS variables are multiple. Indicate, whether to average or to merge results obtained for different variables:

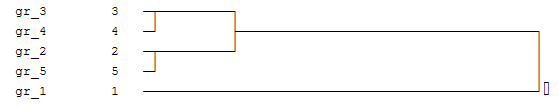
POOL – (default/unspecifying) average the matrices (and hence dendrograms) in one.

CONCAT – append the matrices into the block matrix[[1]](#footnote-1). On a single dendrogram, dendrograms for all the variables will be presented, so one can contrast.

In the CONCAT case, names in the matrix and on the dendrogram are concatenated of the ordinal number of the variable in VARS and (as in else case) the group code. For instance, v03gr\_5 means: the third VARS variable, group coded 5.

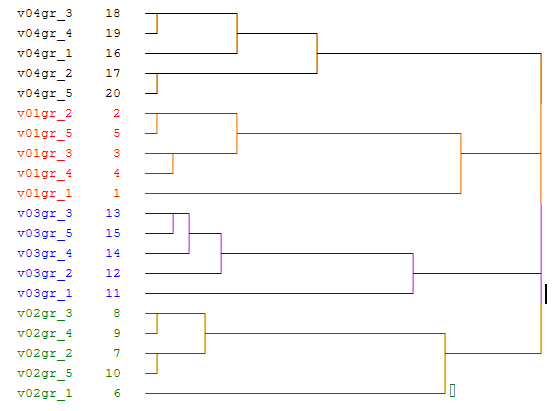
EXAMPLE 1

!KO\_cumclug vars= q13 q14 q15 q16 /grvar= city /measure= CHEB.



* Four variables *Q13, Q14, Q15, Q16* represented data collected in a “price sensitivity” study; their values are the price levels. Variable *CITY* is five cities where the study was conducted (the cities were coded as 1, 2, 3, 4, 5).
* Requested is to measure, with Chebyshev distance, differences between the cities by the shape of cumulative distributions in the variables.
* The distance matrix is saved by the macro as a new dataset, and the dendrogram plotted; it makes visible that, by the 4 variables together, cities 3 and 4, 2 and 5 – are similar, while city 1 is special.

!KO\_cumclug vars= q13 q14 q15 q16 /grvar= city /measure= CHEB /matrxs= CONCAT.



* The same example, but the analysis is separate by the four variables. Because the results are joint in a single dendrogram, one sees that by the 4th variable (v04, i.e. *Q16*) the differences are not as sharp as by the other variables. In the 1st and the 2nd variables the pattern replicates the one obtained on the previous dendrogram.

**MISSING**

If VARS variables are multiple, then how to exclude missing (user- and system-) values in them: listwise, i.e., the case is filtered out when is missing in at least one of the variables (LISTWISE); or from each variable independently (VARIABLE, default/unspecifying the subcommand).

***Limitations***

Maximal width of a value in GRVAR – 5 digits. With MATRXS=CONCAT, maximal width of a value in GRVAR – 2 digits, and maximal number of variables VARS – 99.

***Special regimes***

The macro is not meant for split file state of the input dataset. It obeys case weights, and filtering or selection (commands FILTER, SELECT IF, USE). It obeys temporary (under TEMPORARY) operations.

1. “Empty” cells of such matrix block – lying away from the diagonal – the macro will fill with the largest computed distance increased by one fourth. It is just a filler. [↑](#footnote-ref-1)