UserGuide to the VBA program SRDrepV6T4_CrossValV8D.xlsm

A. General notations

The ***SRDrepV6T4CrossValV8D.xlsm*** file henceforth we call ***SRDrepCrV*** or simply ***program***, the input ***D****ata* ***M****atrix* of the program we denote by ***DM***, the ***N****umber of* ***R****ows,* ***N***umber *of* ***C****olumns,* and ***n****umber of precise* ***dig****its* in ***DM*** we denote by ***nR***, ***nC*** and ***ndig***, respectively.

The ***program*** consists of **10 Excel WorkSheets**: the ***StartSheet*** and the ***pD2, pD3,…, pD8, pD9_15***, ***pDGr15*** probability-distribution (***p-Dist***) Sheets, **8 ModuleSheets** and **one UserForm**. Modules and the UserForm are protected by a password. The ***p-Dist*** Sheets contain the SRD probability distribution tables or functions, actually on ***pD2,…, pD8*** the discrete distribution tables in cases ***nR=2,3,….,8*** for ***ndig=2*** and ***ndig=3***, on ***pD9_15*** and ***pDGr15*** the fitted ***p-Dist*** functions. Sheet ***pD9_15*** contains the parameters of Tangent Hyperbolic (***Tanh***) approximation for ***nR=9,10,…,15***, and ***pDGr15*** contains the parameters of Gaussian (***Normal***) approximation for ***nR>15***. ***Tanh*** and ***Normal*** approximation are there available for ***ndig=2***.

The ***StartSheet*** contains a short description about the inputs, „How to run the program", and results.

The modules contain (altogether) 26 subroutines and 2 functions (this last two for the fitted – ***Tanh*** and ***Normal*** – distribution functions). The only UserForm serves for choosing the ***rPd*** value. The value of ***rPd*** gives for the crossvalidation the ***R***ate of ***P***arts to ***D***elete. For example, if ***nR=16*** and ***rPd=7***, then nR\rPd=16\7=2. It means, that each of the crossvalidation ***minors*** will contain 14 rows from the input ***DM***.

B. Main steps of using *SRDrepCrV*

The figures, helping to understand the main steps, are based on an input file used in food sensory testing (source: [1] L. Sipos, et al.: Journal of Chemometrics, 25 (2011) 275-286.)

1.) Prepare an input ***XLSX*** file containing **an only Excel WorkSheet**, where the cells **B1** and **F1** contain the number of Rows/Columns (***nR/nC***) of ***DM***. The third row contains the names of the Variable vectors and in the **(nC+2)th** column the type of Reference Column (***RC***): Read, min, Max or Average – in the last 3 cases the coordinates of ***RC*** will be evaluated by the program. The first column (starting from 4th row) contains the names of objects – in the example they are abbreviations of sensory properties from [1]. In cells **A2** and **F2** are given an (optional) name of ***DM*** and the (optional) value of ***ndig***. If the user doesn't give the value of ***ndig***, or the value doesn't matches for the available ***p-Dist*** table or function, then the program automatically uses ***ndig=2***, and gives a warning in cells ***E2***, coloring it red and writing into ***E2*** „changed ***ndig***".

	A	B	C	D	E	F	G	H
1	nR= 16				nC= 6			
2	CoExR				ndig= 2			
3		Co1	Co2	Co3	Ex1	Ex2	Ex3	Read
4	YC	51	63	48	32	52	44	45
5	CT	67	68	69	73	75	65	60
6	S	50	73	49	60	59	57	60
7	U	32	56	43	35	24	44	35
8	CS	32	32	43	41	52	46	50
9	SwS	44	42	73	39	32	25	35
10	Scl	42	26	40	51	51	39	45
11	F	67	50	63	64	63	47	60
12	T	43	50	80	58	62	58	60
13	J	45	88	76	55	65	53	60
14	S	20	50	54	64	61	56	60
15	SaF	33	23	7	11	8	17	10
16	SwF	23	65	60	57	48	63	50
17	BoF	50	44	40	47	46	52	50
18	Fl	36	44	58	62	60	63	60
19	Plu	20	61	82	60	61	60	60

Table 1: Input Data

2.) Open the ***SRDrepCrV program***, click on the **START** button, and the program shows the usual ***GetOpenFile*** window, where you have to choose your input file. The program opens the file, and corresponding to ***nR***, either (case-1: ***nR<14***) processes your ***DM*** via **LOO** crossvalidation, or (case-2: **nR>13**) shows a UserForm (named ***StartForm***, see *Fig.1*), where you have to choose an ***rPd*** value.

3.) From the only input WorkSheet will be made the new Sheets by deleting certain rows of the ***DM***, and named the new Sheets **Gr1**x, **Gr2**x, etc, where x is empty in case of **LOO**, and x=**A** or x=**B** else. In case of **LOO** the number of new Sheets is ***nR***, else it is ***nGrA***+***nGrB*** where ***nGrA*** and ***nGrB*** represent the number of Sheets containing a ***minor*** resulted from sequentially (case **A**) or randomly (case **B**) deleted ***rPd*** size part of ***DM***. The value of ***nGrA*** is given by ***nR*** and ***rPd***, the sum of ***nGrA*** and ***nGrB*** equals to the first odd**>50** at a random-cluster-end.

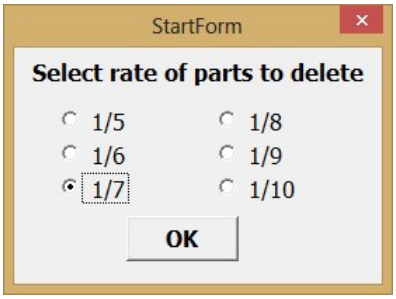

Fig.1: StartForm

4.) The next step of ***SRDrepCrV*** is the wellknown **SRDrep-CRRN** evaluation (including **SRD-Tables**, **CRRN-figures**, and if available, **probability distributions** for the original ***DM's RC*** and for each of the ***minors' RC***.) Here we show only the figures, because they represent clearly the most important data of **SRD-Tables**, as well. The first **CRRN-figure** (*Fig.2*) is the result of the ***DM*** above with the given ***RC***, containing four ties with values 45, 60, 35, and 50, having the lengths 2, 8, 2 and 3, respectively. Because ***nR>15***, for the distribution function ***Normal*** approximation is necessary.

The next figure (*Fig.3*) is the result of a ***minor*** where the (sequentially) deleted 2 rows of ***DM*** were the rows of the objects named **T** and **J**. Because in this case ***nR<=15***, although ***ClassTH=2***, the parameters of the corresponding ***Tanh*** approximation could be found on the ***p-Dist*** Sheet named ***pD9_15***.

About the ***SRDrep%***, ***d***, and ***ClassTH*** parameters you have information in [2].

[2] Kollár-Hunek K., Héberger K.: Classification of SRD-with-Ties probability distributions, Proceeding of Conferentia Chemometrica, Budapest, 13-16 Sept, 2015, **P21**

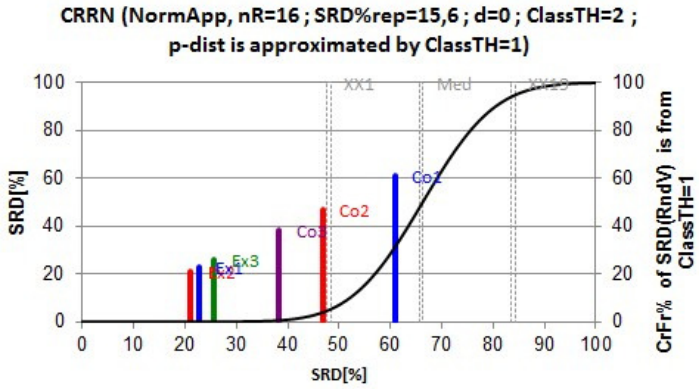

*Fig.2: CRRN figure of the input **DM***

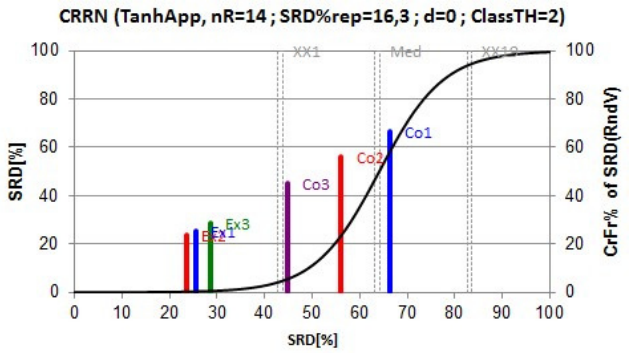

*Fig.3: CRRN figure of a minor. Deleted rows of **DM** are the rows of objects **T** and **J***

	A	B	C	D	E	F	G	H
1	n_row=	16			n_col=	6		
2	CoEx							
3		Co1	Co2	Co3	Ex1	Ex2	Ex3	Read
4	YC	51	63	48	32	52	44	45
5	CT	67	68	69	73	75	65	60
6	S	50	73	49	60	59	57	60
7	U	32	56	43	35	24	44	60
8	CS	32	32	43	41	52	46	60
9	SwS	44	42	73	39	32	25	60
10	Scl	42	26	40	51	51	39	45
11	F	67	50	63	64	63	47	60
12	T	43	50	80	58	62	58	60
13	J	45	88	76	55	65	53	60
14	S	20	50	54	64	61	56	60
15	SaF	33	23	7	11	8	17	10
16	SwF	23	65	60	57	48	63	60
17	BoF	50	44	40	47	46	52	60
18	Fl	36	44	58	62	60	63	60
19	Plu	20	61	82	60	61	60	60

*Table 2: The original **DM** with new **RC***

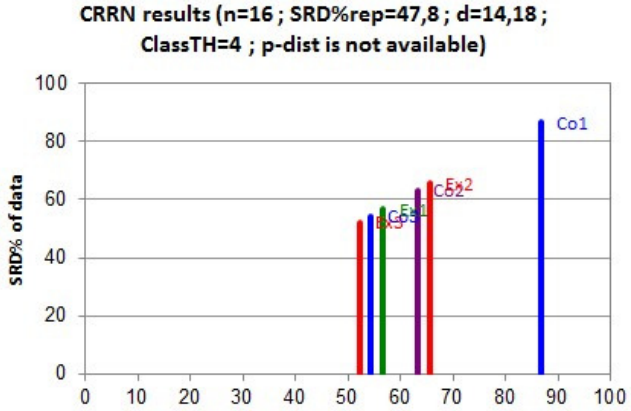

Fig.4: CRRN-figure without p-distribution

This **CRRN-figure** resulted from the same ***DM*** as the first one (*Fig.2*), but now the ***RC*** – as one can see in *Tab.2* – contains a very long tie, so ***ClassTH=4***, that's why ***p-Dist*** is missing

Let us now go back to the first input file, you have seen in *Table 1*. The file contains food profile analysis data, where we compared the efficiency of **co**nsumers (untrained) and **ex**perts (trained) assessors on ***nR=16*** objects (sensory properties of corn). That's why we named the input Sheet ***CoEx_n16***.

The result file of this input contains the following Sheets: the original input Sheet ***CoEx_n16,*** its SRD-CRRN result Sheet: ***Results_CoEx_n16,*** thereafter the ***A-type*** result Sheets renamed from ***Gr1_A, …,Gr8_A***: ***CrV_Gr1_A, …,CrV_Gr8_A***, containing each of them an **A-type** (deleting the ***rPd*** part of ***DM*** sequentially) ***minor*** and its SRDrep_CRRN results. The next group of CrV Sheets consists of the Sheets renamed from ***Gr9_B,…,Gr55_B***: ***CrV_Gr9_B, …,CrV_Gr55_B***, containing each of them a **B-type** (deleting the ***rPd*** part – in this example 2 rows – of ***DM*** randomly) ***minor*** and its SRDrep_CRRN results. The last two (***CV_Tab*** and ***CV_SRDnor_Histo***) Sheets contain „summarizing" tables and figures.

The program collects the ***SRD*** values of the Variable vectors and the ***SRDmax*** value of the corresponding ***RC*** from the ***CrV*** Sheets, and writes them into the ***CV_Tab*** Sheet (see on *Tab. 3*). Thereafter, using the ***SRDmax*** of the corresponding ***RC***, calculates the ***SRDnor*** values (see on *Tab. 4).*

Multiple CV results of type 2-size ordered(A) and 2-size random(B) groups deleted

SRDmax	Group/Var	Co1	Co2	Co3	Ex1	Ex2	Ex3
98	Gr1_A	59	40	39	21	19	27
98	Gr2_A	66	41	37	23	22	25
	***						
96	Gr8_A	54	44	38	22	20	24
96	Gr9_B	56	53	31	22	22	30
96	Gr10_B	62	41	39	24	24	28
	***						
96	Gr54_B	60	55	44	24	23	31
98	Gr55_B	60	47	30	23	22	27

*Table 3: **SRD** values of Variable vectors based on **minors** of **DM***

CV(AB) results given by SRDnor(=100*SRD/SRDmax)

SRDmax	Group/Var	Co1	Co2	Co3	Ex1	Ex2	Ex3
98	Gr1_A	60.2	40.8	39.8	21.4	19.4	27.6
98	Gr2_A	67.3	41.8	37.8	23.5	22.4	25.5
	***						
96	Gr8_A	56.3	45.8	39.6	22.9	20.8	25.0
96	Gr9_B	58.3	55.2	32.3	22.9	22.9	31.3
96	Gr10_B	64.6	42.7	40.6	25.0	25.0	29.2
	***						
96	Gr54_B	62.5	57.3	45.8	25.0	24.0	32.3
98	Gr55_B	61.2	48.0	30.6	23.5	22.4	27.6

*Table 4: **SRDnor** Values of Variable vectors based on **minors** of **DM***

CV(A) results given by ordered SRDnor

OrdNum	Co1	Co2	Co3	Ex1	Ex2	Ex3
1	56.3	40.8	31.3	20.4	19.4	20.8
2	58.3	41.8	36.5	20.8	19.4	24.5
3	59.2	45.8	37.8	21.4	19.8	25.0
4	60.2	45.8	39.6	22.9	19.8	25.0
5	60.4	46.9	39.8	23.5	20.8	25.5
6	64.3	49.0	39.8	24.0	22.4	27.6
7	66.7	52.0	41.7	25.0	23.5	28.6
8	67.3	53.1	42.9	25.5	25.0	31.3
Min	56.3	40.8	31.3	20.4	19.4	20.8
Med	60.3	46.4	39.7	23.2	20.3	25.3
Max	67.3	53.1	42.9	25.5	25.0	31.3
Average	61.6	46.9	38.6	22.9	21.3	26.0
StDev	3.8	4.1	3.4	1.8	2.0	2.9

*Table 5: Ordered SRDnor values from the **A-type** **CrV_Gr1_A**,---, **CrV_Gr8_A** Sheets*

In *Tab.3* and *Tab.4* the ***SRD*** or ***SRDnor*** values are ordered by the crossvalidation WorkSheets. The normalization of ***SRD*** values by the ***SRDmax*** value, belonging to the ***RC*** of the Sheet, allows the monotonic increasing ordering of ***SRDnor*** values. *Tab.5* shows these ordered values of the **A-type** part of *Tab.4*, completed by minimum, median, maximum, average and standard deviation, column by column.

Tab.6, evaluated from *Tab.5*, contains the frequencies and cumulative frequencies of the ***SRDnor*** data sets.

All of the *Tables 3* to *6* are on Sheet ***CV_Tab***. The last WorkSheet of the program file, named ***CV_SRDnor_Histo***, consists of similar tables as *Tab.5* and *Tab.6*, but now for the whole data set of **A-type** and ***B-type minors'*** ***SRDnor*** values. The **Histogram CV(AB)** table's structure is the same, as **Histogram CV(A)**'s, but it contains the histogram-data of the whole (ordered) *Tab 4*, so one can easily imagine it.

Histogram CV(A)

	Co1			Co2			Co3			Ex1			Ex2			Ex3	
SRDnor	freq	Cfreq	SRDnor	freq	Cfreq	SRDnor	freq	Cfreq	SRDnor	freq	Cfreq	SRDnor	freq	Cfreq	SRDnor	freq	Cfreq
56.3	1	1	40.8	1	1	31.3	1	1	20.4	1	1	19.4	2	2	20.8	1	1
58.3	1	2	41.8	1	2	36.5	1	2	20.8	1	2	19.8	2	4	24.5	1	2
59.2	1	3	45.8	2	4	37.8	1	3	21.4	1	3	20.8	1	5	25.0	2	4
60.2	1	4	46.9	1	5	39.6	1	4	22.9	1	4	22.4	1	6	25.5	1	5
60.4	1	5	49.0	1	6	39.8	2	6	23.5	1	5	23.5	1	7	27.6	1	6
64.3	1	6	52.0	1	7	41.7	1	7	24.0	1	6	25.0	1	8	28.6	1	7
66.7	1	7	53.1	1	8	42.9	1	8	25.0	1	7				31.3	1	8
67.3	1	8							25.5	1	8						

*Table 6: Frequency and Cumulative frequency of SRDnor values by Variable vectors in the **A-type** crossvalidation minors*

Ordered CV(AB) columns given by SRDnor(=100*SRD/SRDmax)						
OrdNum	Co1	Co2	Co3	Ex1	Ex2	Ex3
1	52.0	38.8	27.1	20.4	18.4	19.4
2	52.1	39.8	30.6	20.4	18.4	19.4
***						
54	70.8	56.1	45.8	26.0	25.0	33.3
55	70.8	57.3	45.8	26.0	25.0	33.3
Min	52.0	38.8	27.1	20.4	18.4	19.4
Q1	58.3	42.7	36.7	21.4	19.8	25.0
Med	60.4	46.9	38.8	22.9	20.8	25.5
Q3	64.6	51.0	40.6	23.5	22.4	28.6
Max	70.8	57.3	45.8	26.0	25.0	33.3
Average	61.4	47.2	38.7	22.8	21.2	26.3
StDev	4.6	4.8	3.9	1.5	1.9	3.3
Bin Averages for Min(0;1/8), Q1(1/8;3/8), Med(3/8;5/8), Q3(5/8;7/8), Max(7/8;1)						
Min	53.8	40.0	30.7	20.6	18.4	20.2
Q1	58.1	43.0	36.8	21.6	19.6	24.4
Med	60.7	47.1	38.7	22.7	21.0	25.8
Q3	64.3	50.5	40.8	23.6	22.7	28.2
Max	69.1	54.8	44.6	25.5	24.2	31.8

*Table 7: Ordered SRDnor values with statistical attributes from all of the **CrV_Gr** Sheets*

The prior described *Tab.5* is similar to *Tab.7*, but the second one contains the ordered SRDnor data for all of the ***CrV_Gr*** Sheets, wich data set is big enough to investgate ***Q1*** and ***Q3*** percentiles, as well. As an additional part, here we see the *Bin Averages* belonging to the three given percentiles. These data helped us to create a very expressive report of the crossvalidation results, which we placed on the second Worksheet, named in our example ***Results_CoEx_n16***. This report consists of *Tab.8* and *Fig.5.* In *Tab.8* the first row lists the ordered ***SRDnor*** data of the original input ***DM***. It is interesting to compare them with the values of crossvalidation results, namely with the median, bin average of median and total average. Considering the example we see the best conformity in case of the Variable vectors Ex2 and Ex1, but the others are not very much worse, either. The Box-Whiskers figure (**BWfig**) strenghtens the ***SRD-CRRN*** results, as well.

As last part of this section we have to mention that neither *Tab.7* and *Tab.8*, nor the *Box-Whishkers* figures are available for LOO crossvalidation, where ***nR<14*** in ***DM***. It means that the number of crossvalidation ***minors*** is <14, as well. In this case for the characterisation of the Variable vectors is enough to consider their Min, Med and Max values.

Histogram bins by percentiles, V-columns ordered by the Median of the corresponding CrV histogram															
SRDnor	21.1	Ex2		22.7	Ex1		25.8	Ex3		38.3	Co3		46.9	Co2		60.9	Co1	
	SRDnor	BinAve		SRDnor	BinAve		SRDnor	BinAve		SRDnor	BinAve		SRDnor	BinAve		SRDnor	BinAve	
Min(0;1/8)	18.4	18.4		20.4	20.6		19.4	20.2		27.1	30.7		38.8	40.0		52.0	53.8	
Q1(1/8;3/8)	19.8	19.6		21.4	21.6		25.0	24.4		36.7	36.8		42.7	43.0		58.3	58.1	
Med(3/8;5/8)	20.8	21.0		22.9	22.7		25.5	25.8		38.8	38.7		46.9	47.1		60.4	60.7	
Q3(5/8;7/8)	22.4	22.7		23.5	23.6		28.6	28.2		40.6	40.8		51.0	50.5		64.6	64.3	
Max(7/8;1)	25.0	24.2		26.0	25.5		33.3	31.8		45.8	44.6		57.3	54.8		70.8	69.1	
Total Ave=		21.2	Total Ave=		22.8	Total Ave=		26.3	Total Ave=		38.7	Total Ave=		47.2	Total Ave=		61.4	
Ex2 StDev=		1.9	Ex1 StDev=		1.5	Ex3 StDev=		3.3	Co3 StDev=		3.9	Co2 StDev=		4.8	Co1 StDev=		4.6	

Table 8: Description of Variable vectors by SRDnor and CrossValidation Statistics

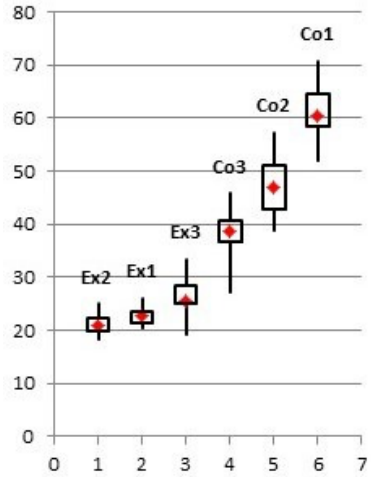

Fig.5. Box-Whiskers figures

C. About the bounds of input data

Among the input data ***nR*** and ***nC*** are strictly bounded by the program, namely ***4≤nR≤1300*** and ***2≤nC≤300***. Of course, with increasing ***nR*** or ***nC*** the running time of the ***program*** quickly increases. In case of increasing ***nC*** one have to take in account the Excel-bound for the labels of data series, too. In case of equal ***SRD*** values the names of V-vectors on the **CRRN figure** are concatenated – and if the length of a concatenated name is >100, then the ***program*** writes on the figure only the cells of the whole name. Similarly, there is an Excel-bound for the number of data-series (***nDs***) on a figure. That's why for ***nC>63*** (where ***nDs=4*nC≥256***), **BWfigs** are segmented by the ***program***. For example, if ***nC=137*** (and ***nR***>13) then the result file of crossvalidation contains 1+136\60=**3 BWfig** segments, containing the first and second ones 46 and the third one 45 V-vectors' **BWfig**.