Quick reference for CKM in τ-Argus 4.2.0

Date: June 2020

Author: Peter-Paul de Wolf

Abstract:

This Quick Reference is an addition to the manual of τ-ARGUS version 4.1, describing the new functionality in τ-ARGUS version 4.2 that can be used to add noise to tables using the Cell Key Method.

Content

Introduction and background

Since version 4.2.0 a new protection method is included in τ-ARGUS: noise addition using the Cell Key Method (CKM for short). For a description of this method, see deliverables 4.1 and 4.2 of the SGA "Open Source tools for perturbative confidentiality methods", partly funded by Eurostat (Contract N° 2018.0108 under FPA N° 11112.2014.005-2014.533).

The cell key method as we implemented it, is based on the method proposed in Fraser and Wooton, 2005 (see also Marley and Leaver (2011) and Thompson et al. (2013)) by the Australian Bureau of Statistics (ABS). Taking up on suggestions for a special extension suggested in Ma et al. (2016) on the way.

The method can be applied to frequency count tables as well as magnitude tables.

In the subdirectory `data/cellkey/Frequency` of the τ-ARGUS installation, two test datasets are provided:

1. the synthetic dataset `hc_9_2_synth_Tau.csv` in **free** format, based on the file a file as given in the zipfile on https://ec.europa.eu/eurostat/cros/content/3-random-noise-cell-key-method_en. This is the test synthetic dataset used in the previous SGA "Harmonised protection of Census data", containing 820000 records;
2. the dataset `tau_testW_CKM.asc` based on the "standard" τ-Argus test dataset. This microdata set is in **fixed** format.

To be able to apply CKM, recordkeys must be present in the dataset and a ptable-file must be provided. For more details, see the section Prerequisites.

In case of applying CKM to magnitude tables, the parameters are specified in the `.rda` file and can be adjusted in the GUI (Specify Metadata).

References

Fraser, B. and Wooton, J. (2006). *A proposed method for confidentialising tabular output to protect against differencing*. In: Monographs of Official Statistics. Work session on Statistical Data Confi-dentiality, Eurostat-Office for Official Publications of the European Communities, Luxembourg, 2006, pp. 299-302.

Ma Y., Lin YX., Chipperfield J., Newman J., Leaver V. (2016) A New Algorithm for Protecting Aggregate Business Microdata via a Remote System. In: Domingo-Ferrer J., Pejić-Bach M. (eds) Privacy in Statistical Databases. PSD 2016. Lecture Notes in Computer Science, vol 9867. Springer, Cham

Marley, J. K. and Leaver, V. L.: A method for confidentialising user-defined tables: Statistical proper-ties and a risk-utility analysis. In: Proceedings of 58th World Statistical Congress, pages 1072– 1081 (2011)

Thompson, G., Broadfoot, S., Elazar, D. (2013): "Methodology for the Automatic Confidentialisation of Statistical Outputs from Remote Servers at the Australian Bureau of Statistics", paper pre-sented at the Joint UNECE/Eurostat Work Session on Statistical Data Confidentiality (Ottawa, 28- 30 Oktober 2013) available at http://www.unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.46/2013/Topic_1_ABS.pdf

Prerequisites

To be able to apply CKM, one needs the following:

1. Record keys should be available in the microdata
2. One or more p-table files should be available

Record keys

The way we have implemented CKM differs slightly from the "original" ABS method. What differs most, is the representation of the record keys. In our implementation the record keys should be independent uniformly distributed values between zero and one.

To produce record keys, one could use e.g. the following R-script (free format microdata):

```
# Record keys

microdata <- read.csv2(file="hc_9_2_synth.csv")
set.seed(6197257)
microData$recordkey <- format(runif(dim(microData)[1]), scientific=FALSE, nsmall=15)
write.table(microData, "hc_9_2_synth_Tau.csv", row.names=FALSE, col.names=FALSE, quote=FALSE,
dec=".",sep=";")
```

This script adds a record key variable with 15 decimals to an existing (free format) microdata file. The number of decimals needs to be specified in the metadata file (`.rda`) and should be sufficient to be able to distinguish between the different probabilities in the p-table file(s).

To benefit from the CKM method, it is essential that the same record keys are used when protecting tables based on the same microdata!

Only one record key variable per microdata file is allowed. Note that this implies that the same record key is thus used for application of CKM to any of the variables within the same microdata file.

A record key variable is specified in the `.rda` file as follows:

```
record_key 17
       <RECORDKEY>
       <DECIMALS> 15
       <PFILE_FREQ> "ptab_freq_all.csv"
       <PFILE_CONT> "ptab_savings_p1_all.csv"
       <PFILE_SEP>  "ptab_savings_p1_smallcells.csv"
```

The tag `<PFILE_FREQ>` is used to specify the name of the p-table file to be used for frequency count tables, the tag `<PFILE_CONT>` to specify the name of the p-table file to be used for magnitude tables and `<PFILE_SEP>` to specify the name of the p-table file to be used for small cell values in magnitude tables (when using the "separation" option).

p-table file(s)

For a detailed description of the ideas behind a p-table, see deliverables D4.1 and D4.2 of the SGA "Open Source tools for perturbative confidentiality methods", partly funded by Eurostat (Contract N° 2018.0108 under FPA N° 11112.2014.005-2014.533).

p-table files can be generated by the R-package ptable (see https://github.com/sdcTools/ptable) or can be constructed by yourself. When using the ptable package (version 0.3.3 or higher) save the p-table in `destatis` format for use in τ-ARGUS. P-table files are simple ;-separated ASCII files.

The format of a p-table file for frequency count tables differs slightly from that for magnitude tables. However, the general structure is that the first line of the file contains column labels and is followed

by lines with values. The following subsections will show the details for frequency count tables and magnitude tables respectively.

p-table for frequency count tables

For frequency count variables the variables should be (in this order):

`i`, `j`, `p_ij`, `v` and `p_ij_ub`

where `i` is the true cell value, `j` is the value to change into, `p_ij` is the probability to change `i` into `j`, `v` is the difference `j-i` and `p_ij_ub` is the cumulative probability. The values for `i` and `j` should be integer. The labels that are used are free to choose.

Example (first few lines of a p-table file for frequency count tables):

```
i;j;p_ij;v;p_ij_ub
0;0;1;0;1
1;0;0.666749933;-1;0.666749933
1;3;0.333000567;2;0.9997505
1;4;0.0002493;3;0.9999998
1;5;1.00E-07;4;0.9999999
1;6;1.00E-07;5;1
2;0;0.3334165;-2;0.3334165
2;3;0.6663346;1;0.9997511
2;4;0.0002486;2;0.9999997
2;5;1.00E-07;3;0.9999998
2;6;1.00E-07;4;0.9999999
2;7;1.00E-07;5;1
...
```

p-table for magnitude tables

For frequency count variables the variables should be (in this order):

`i`, `j`, `p_ij`, `v`, `p_ij_ub` and `type`

where `i` is the true cell value, `j` is the value to change into, `p_ij` is the probability to change `i` into `j`, `v` is the difference `j-i`, `p_ij_ub` is the cumulative probability and `type` is `all`, `even` or `odd`. Type determines whether a p-table is used for cells with any number of contributions, even number of contributions or odd number of contributions respectively. The values for `i` should be integer, the values for `j` can be doubles. The labels that are used are free to choose.

Example (first few lines of a p-table file for magnitude tables):

```
i;j;p;v;p_int_ub;type
0;0;1;0;1;all
1;0;0.24006614;-1;0.24006614;all
1;0.5;0.24006614;-0.5;0.48013228;all
1;1;0.24006615;0;0.72019843;all
1;1.5;0.10377025;0.5;0.82396868;all
1;2;0.06682983;1;0.89079851;all
1;2.5;0.04250929;1.5;0.9333078;all
1;3;0.02670627;2;0.96001407;all
1;3.5;0.01657137;2.5;0.97658544;all
1;4;0.01015593;3;0.98674137;all
1;4.5;0.00614748;3.5;0.99288885;all
1;5;0.00367527;4;0.99656412;all
1;5.5;0.0021702;4.5;0.99873432;all
1;6;0.00126568;5;1;all
...
```

Batch file

First of all, the <WRITETABLE> command has the option added to save the table in CKM-format. The CKM format is similar to the Code-Value format: first the spanning variables are given, then the perturbed cell value. Optionally, the original cell value, the difference and the cell key can be added.

Each line in the output file would thus look like

Var1; Var2; …; VarX; Cell Value*; Original Cell Value; Difference; Cell-Key*

An example line in the output would be

Total; Nr; 11393*; 11395; –2; 0.204963078*

where the italic values are indicating the optional values.

To use the CKM-format in the <WRITETABLE> command, specify output-type 7. Options can be turned on (default is off) using AO (add original) AD (add difference) and AC (add cell key).

Example:

```
<WRITETABLE>      (1, 7, AO+AD-AC+, "C:\Argus\data\cellkey\output.tab")
```

would result in an CKM-format output file with additionally the original values and the cell key, but not the difference. The previous example would thus yield a line like

Total; Nr; 11393; 11395; 0.204963078

Secondly, it is possible to apply CKM using the τ-ARGUS batch-file approach (using the .arb file). The <SUPPRESS> tag can now be used to apply CKM:

```
<SUPPRESS> CKM(TabNo)
```

With TabNo the number of the table to be protected (similar as with other protection methods). In this case the p-table file as specified in the corresponding metadata file (.rda) is used.

Additional option for frequency count tables

For frequency count tables you can optionally specify an alternative p-table file to be used. The specified p-table file would then replace the one (if any) that is specified in the corresponding metadata file (.rda) under the <PTABLE_FREQ> tag.

Examples:

```
<SUPPRESS>  CKM(1)

<SUPPRESS>  CKM(2) "C:\Argus\Ptables\ptable4.csv"
```

Additional options for magnitude tables

For magnitude tables, three optional arguments are available: alternative filename for PTABLE_CONT, alternative filename for PTABLE_SEP and value for additional perturbance for unsafe cells (μ_C). The arguments are separated by |. All the vertical bars should be present whenever at least one optional argument is used

Examples:

```
<SUPPRESS>  CKM(1)
<SUPPRESS>  CKM(2) "C:\Argus\Ptables\altptableCONT.csv" ||
<SUPPRESS>  CKM(3) | "C:\Argus\Ptables\altptable_SEP.csv" |
<SUPPRESS>  CKM(4) "C:\Argus\Ptables\altptable_CONT.csv" || 2.7
<SUPPRESS>  CKM(5) || 2.7
<SUPPRESS>  CKM(6) "altptable_CONT.csv" | "altptable_SEP.csv" | 2.7
```

Specifying methods/parameters for magnitude tables in the `.rda`-file

Applying CKM to magnitude tables is available through specification in the `.rda`-file or through the GUI (see the walk through for more information on the GUI). For specification in the `.rda`-file several new keywords are introduced, see Table 1. For explanation of the different options, we refer to deliverable D4.2 Part I of the SGA "Open Source tools for perturbative confidentiality methods", partly funded by Eurostat (Contract N° 2018.0108 under FPA N° 11112.2014.005-2014.533).

Keyword	Description	Values
<CKM>	Defines the variant of CKM to be used in case of magnitude table. N is not allowing CKM, T is using TopK (=integer) observations, M is using cell mean, D is using distance between largest and smallest observation and V is using the cell value (see section 2.1 in deliverable D4.2).	- N - T(TopK) - M - D - V *Default: N* *Restriction: 1 ≤ TopK ≤ 5*
<INCLUDEZEROS>	Defines if zero contributions should be included (Y) or not (N) in the calculation of cell keys (see section 2.2 in deliverable D4.2).	- Y - N *Default: N*
<PARITY>	Defines if even and odd counts shall be perturbed differently. If "Y" the ptable file should contain probabilities for even AND odd (see section 3.2 in D4.2).	- Y - N *Default: N* *Restriction: if* **<CKM>** *T(TopK) with TopK > 1 is used, only* **<PARITY>** *N is allowed*
<SCALING>	"F" defines that the flex function shall be used. "N" defines not to use the flexfunction, but then still some parameters need to be	- F(σ_0, σ_1, x^*, q, ϵ_2, ϵ_3, …, ϵ_T) - N(σ_1, ϵ_2, ϵ_3, …, ϵ_T) *No default: mandatory when using* **<CKM>**

	provided. See section 2.3 in deliverable D4.2.	
<SEPARATION>	Defines if small values shall be treated like counts.	- Y - N *Default: N*
<PFILE_FREQ> or <PFILE>	Sets the path to a ptable file to be used for count tables	- String with file name
<PFILE_CONT>	Sets the path to a ptable file to be used for magnitude tables	- String with file name
<PFILE_SEP>	Sets the path to a ptable file to be used when <SEPARATION> = Y	- String with file name

Table 1: Commands/keywords to be used in the .rda-file to specify the way CKM should be used

The keywords <CKM>, <INCLUDEZEROS>, <PARITY>, <SCALING> and <SEPARATION> need to be specified for each numeric variable. Whenever a numeric variable has *no* <CKM> tag, the default is taken, i.e., CKM is not allowed for that variable.

The keywords <PFILE_FREQ>, <PFILE_CONT> and <PFILE_SEP> are connected to the <RECORDKEY> variable and are thus specified once and used for all numeric variables and for frequency count tables.

Additionally, the first implementation has some settings suggested in deliverable D4.2 part I fixed or not implemented at all. These are:

- Same_Key: the perturbation of the cellkey is *always* applied (section 2.2 of D4.2)
- Special treatment for positive and negative contributions (section 2.6 of D4.2): only option 2 is implemented.
- Mean before sum is not implemented (section 2.7 of D4.2).

Finally, the additional protection for sensitive cells (section 2.4 of D4.2 part I) can be set through the GUI. If the table has sensitive cells identified and CKM is applied, a dialog will pop up to specify the extra amount of perturbation. Setting that value to 0.0 means that *no* additional perturbation will be used.

Example rda-file:

```
<SEPARATOR> ";"
Sex 1
  <RECODEABLE>
  <HIERARCHICAL>
  <HIERCODELIST> "dim_sex.hrc"
  <HIERLEADSTRING> "@"
Age 1
  <RECODEABLE>
  <HIERARCHICAL>
  <HIERCODELIST> "dim_age.hrc"
  <HIERLEADSTRING> "@"
HighIncome 1
  <NUMERIC>
Savings 4
  <NUMERIC>
```

```
  <CKM> T(3)
  <INCLUDEZEROS> N
  <SEPARATION> Y
  <PARITY> N
  <SCALING> F(0.03,0.30,1000,3,0.5,0.2)
Weigt 3
  <WEIGHT>
Key 10
  <RECORDKEY>
  <DECIMALS> 8
  <PFILE_FREQ> "ptab_freq_total.csv"
  <PFILE_CONT> "ptab_savings_p1_all.csv"
  <PFILE_SEP>  "ptab_savings_p1_smallcells.csv"
```

For CKM applied to magnitude tables, we suggest to use record keys with at most 8 decimals. Similarly, in the p-table the probabilities a recommended to have no more than 8 digits.

Information loss measures

Currently, only in case of CKM applied to frequency count tables some information loss measures are provided by τ-ARGUS.

The distribution of the realized differences is provided both visually in the protected table as well as in the table summary. Visually: darker color means greater difference. Table summary: number and percentage of cells for each value of the occurring differences (empty cells are counted separately).

The following information loss measures are calculated:

- Absolute Difference $|D_{pert}(c) - D_{orig}(c)|$
- Relative Absolute Difference $\left(D_{pert}(c) - D_{orig}(c)\right)/D_{orig}(c)$
- Absolute Difference of square Roots $|\sqrt{D_{pert}(c)} - \sqrt{D_{orig}(c)}|$

where $D_{\dots}(c)$ stands for the value of cell c and c runs over all cells in the table.

For all distances descriptives are given (minimum, median, mean, maximum) as well as percentiles (P10, P20, P30, P40, P50, P60, P70, P80, P90, P95, P99).

Information related to the Empirical Cumulative Distribution Functions (ECDF) of all distances is given:

- The number and the fraction of cells with an Absolute Difference less than or equal to 0, 1, 2, 3, 4, …, maximum perturbation, Infinity.
- The number and the fraction of cells with a Relative Absolute Difference less than or equal to 0, 0.02, 0.05, 0.10, 0.20, 0.30, 0.40, 0.50, 1.00, Infinity.
- The number and the fraction of cells with an Absolute Difference of square Roots less than or equal to 0, 0.02, 0.05, 0.10, 0.20, 0.30, 0.40, 0.50, 1.00, Infinity.
- For all distances a graph can be displayed of the Empirical Cumulative Distribution Function.

Number of false non-zero cells is given (original value zero changed into a non-zero).

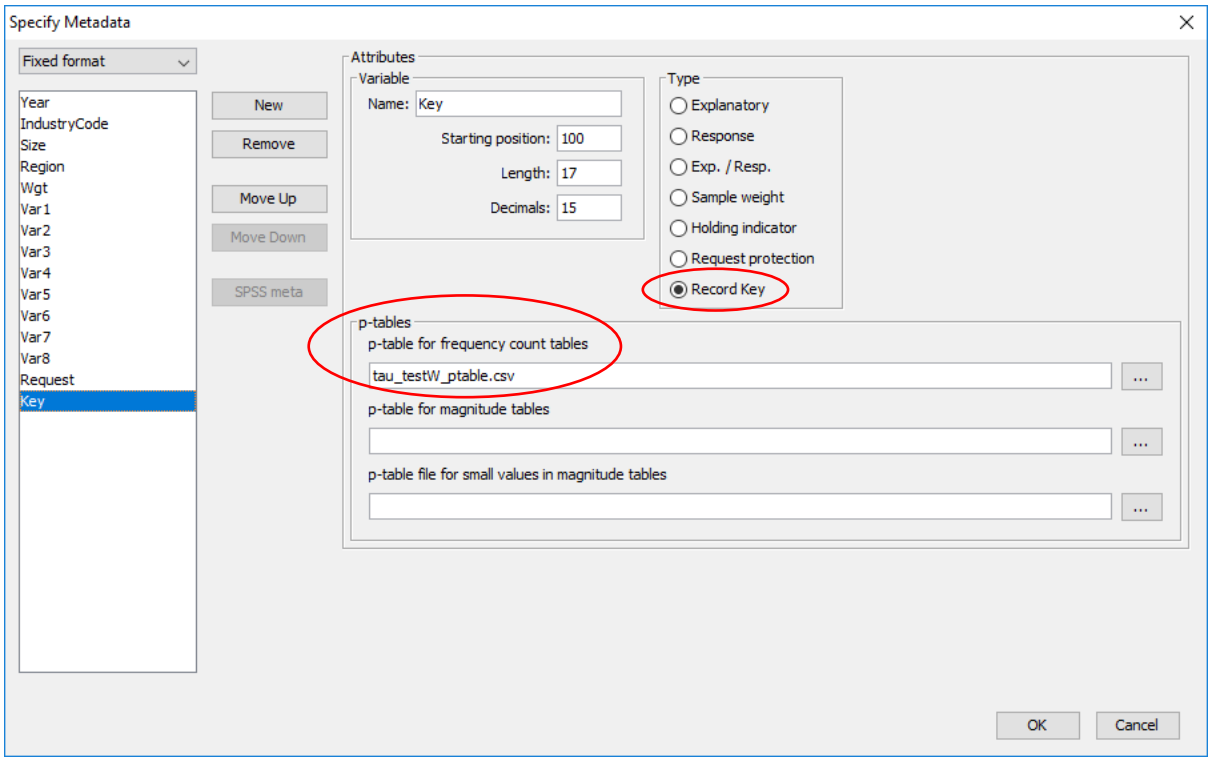
Number of false zero cells is given (original value non-zero changed into zero).

Known issues/restrictions

- Record keys need to be a separate variable in the microdata.
- In free format, τ-Argus can only deal with non-scientific notation of numeric variables. This is a known general τ-Argus issue.
- Numeric variables (such as the probabilities and the record keys) should have a dot as decimal separator.
- The p-table must be in Destatis-format (use the argument `type="destatis"` when creating the p-table from within the R-package ptable).
- Holdings cannot be taken into account.
- The manual is not yet updated.
- As in general with τ-Argus, reading fixed format microdata input is (much) faster compared to reading free format.
- For magnitude tables, only a first implementation is provided, with no information loss measures nor table summary.

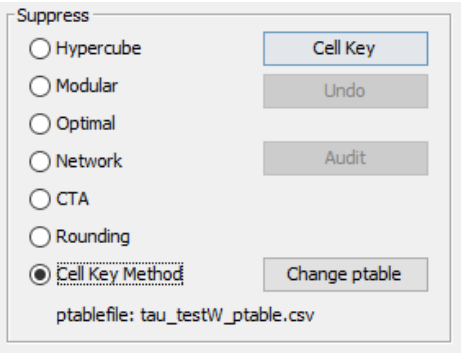
CKM applied to frequency count variables, a walk through

1. Open a microdata file and specify the corresponding metadata (e.g. via the `.rda`-file). In the next screenshot of 'Specify Metadata' you can see how the record-key is selected and the corresponding p-table file is selected. In this case, only a p-table file for applying CKM to frequency count tables is specified.



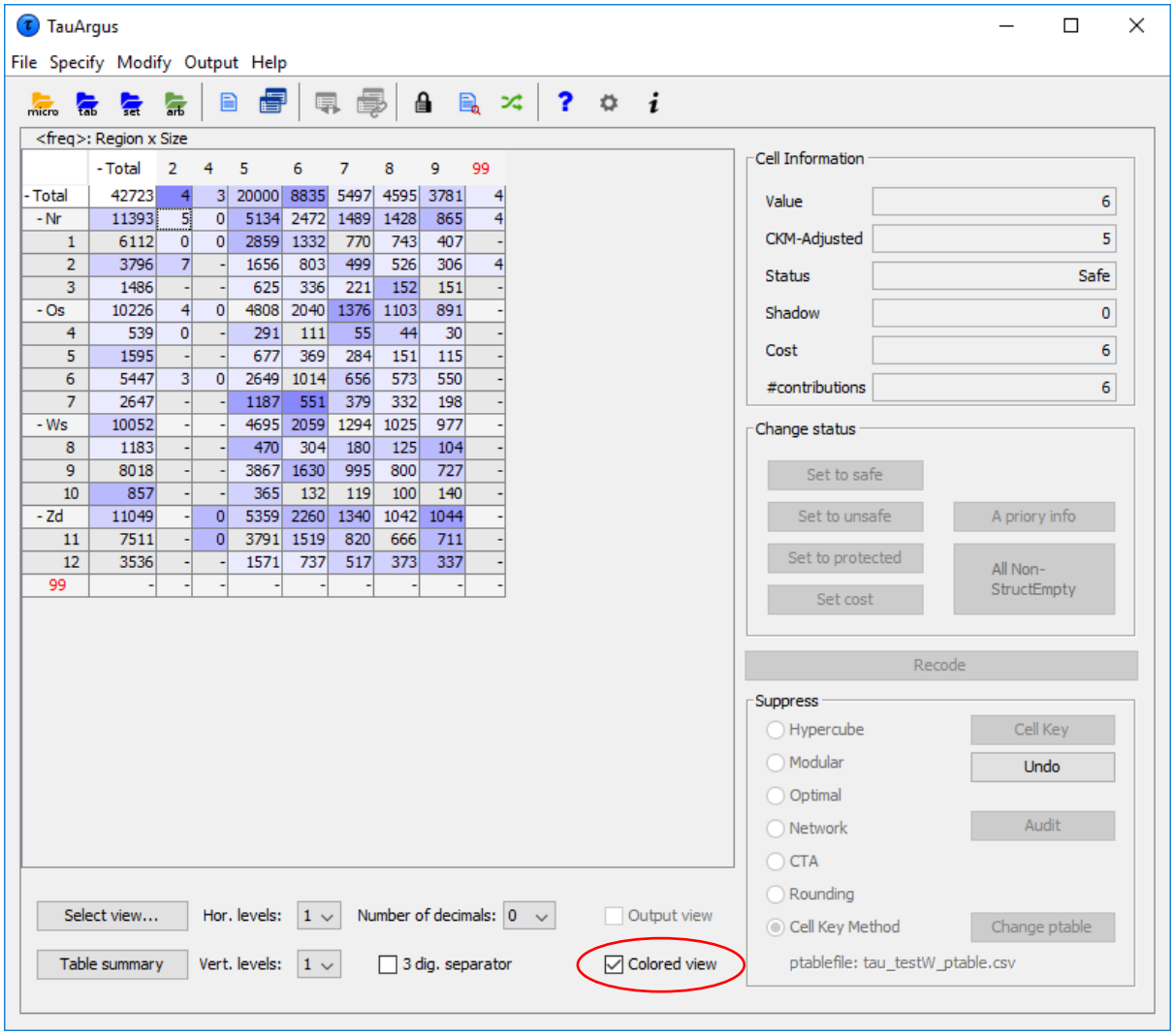
2. Specify the table to which you want to apply the cell key method. When specifying the table, you need to select `<freq>` as response variable. Only then the cell key method will become available.
   Since the cell-key method is a method that will be applied to all cells in the same way irrespective of the fact whether they are unsafe or not, it does not make sense to specify any safety rules when specifying the table. It is allowed though.

3. Select the Cell Key Method in the Suppress-panel on the main screen:



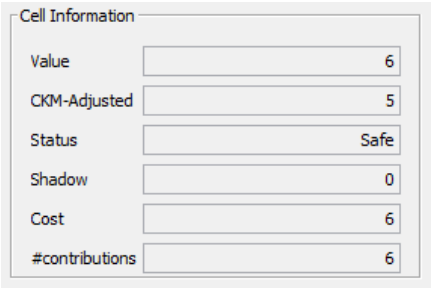
If you want you can choose a different p-table file by pressing the ‘Change ptable’ button.

4. To apply the Cell Key Method, press the button labelled Cell Key.
5. You will get visual feedback by graded colors (darkest = largest amount of noise added, no-color = no noise added) by checking the ‘Colored view’ check-box:

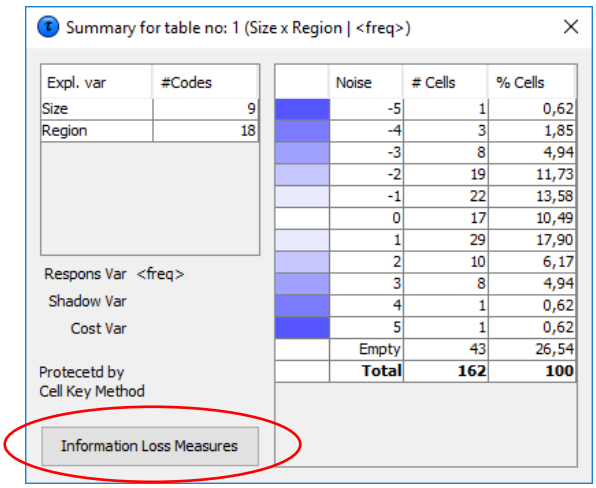


	- Total	2	4	5	6	7	8	9	99
- Total	42723	4	3	20000	8835	5497	4595	3781	4
- Nr	11393	5	0	5134	2472	1489	1428	865	4
1	6112	0	0	2859	1332	770	743	407	-
2	3796	7	-	1656	803	499	526	306	4
3	1486	-	-	625	336	221	152	151	-
- Os	10226	4	0	4808	2040	1376	1103	891	-
4	539	0	-	291	111	55	44	30	-
5	1595	-	-	677	369	284	151	115	-
6	5447	3	0	2649	1014	656	573	550	-
7	2647	-	-	1187	551	379	332	198	-
- Ws	10052	-	-	4695	2059	1294	1025	977	-
8	1183	-	-	470	304	180	125	104	-
9	8018	-	-	3867	1630	995	800	727	-
10	857	-	-	365	132	119	100	140	-
- Zd	11049	-	0	5359	2260	1340	1042	1044	-
11	7511	-	0	3791	1519	820	666	711	-
12	3536	-	-	1571	737	517	373	337	-
99	-	-	-	-	-	-	-	-	-

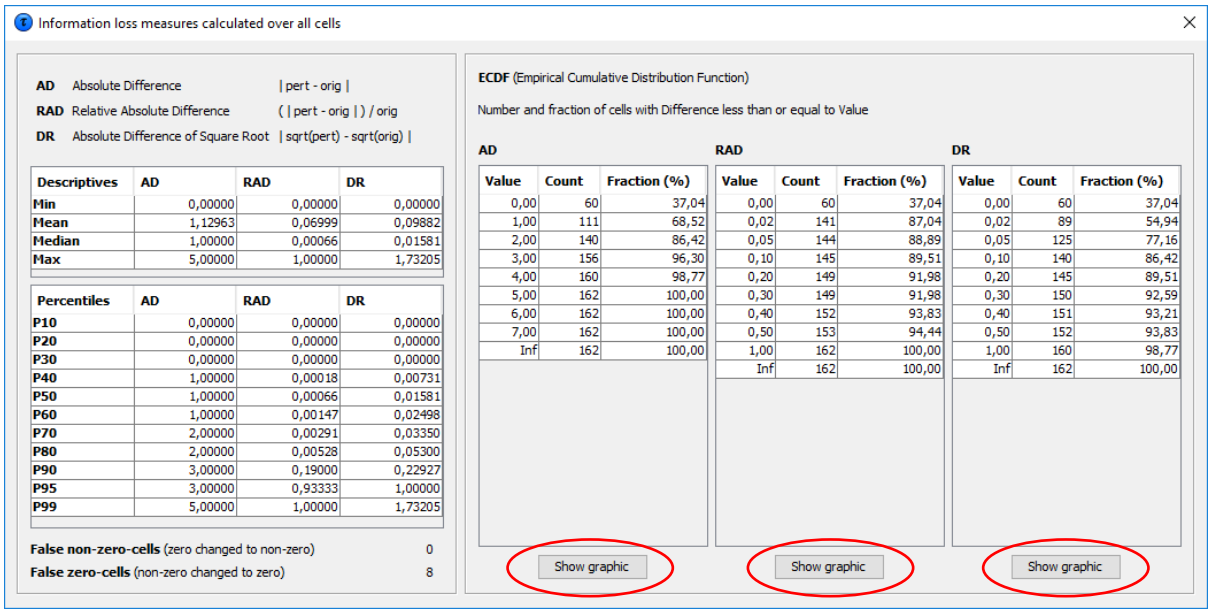
6. Additionally, you can get information on each cell by clicking on that cell:



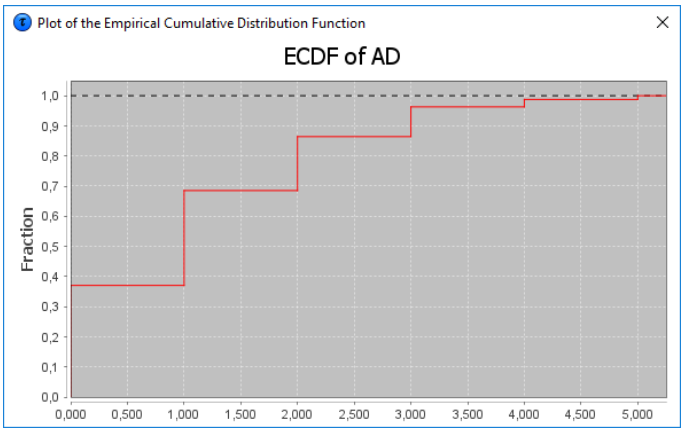
You can get summary information via the button ‘Table Summary’:



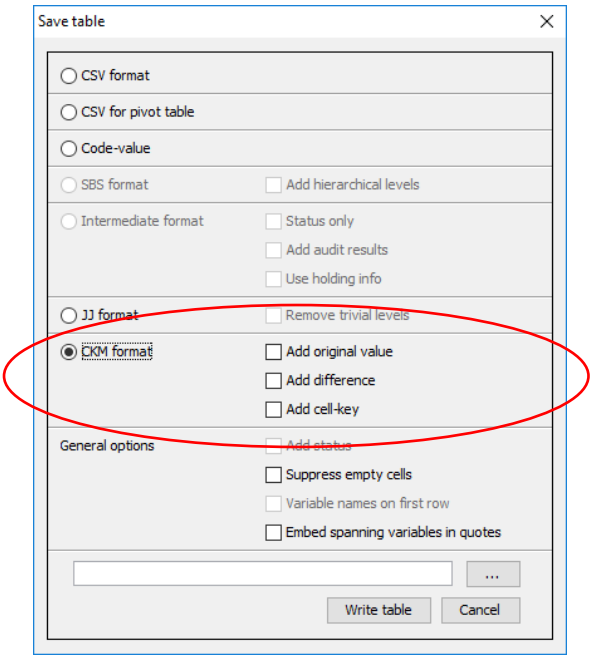
Additional information loss measures by clicking the button ‘Information Loss Measures’:



A graphical representation of the full ECDF of a particular distance can be obtained by clicking the corresponding button ‘Show graphic’:



7. You can save the protected table in the usual formats (except `Intermediate format' and `SBS format'). Additionally, you can use a new format:



This will result in a `.tab` file where each line will contain (Italics means 'as far as selected', i.e., is optional):

Var1; Var2; …; VarX; Cell Value*; Original Cell Value; Difference; Cell-Key*

So e.g.,

Total; Nr; 11393*; 11395; –2; 0.204963078*

8. The report file will contain information on the applied method as well as the information loss measures calculated over all non-empty cells as well as over all cells (including empty cells). Note that this primarily affects the percentages, see e.g., the summary table in the report file:

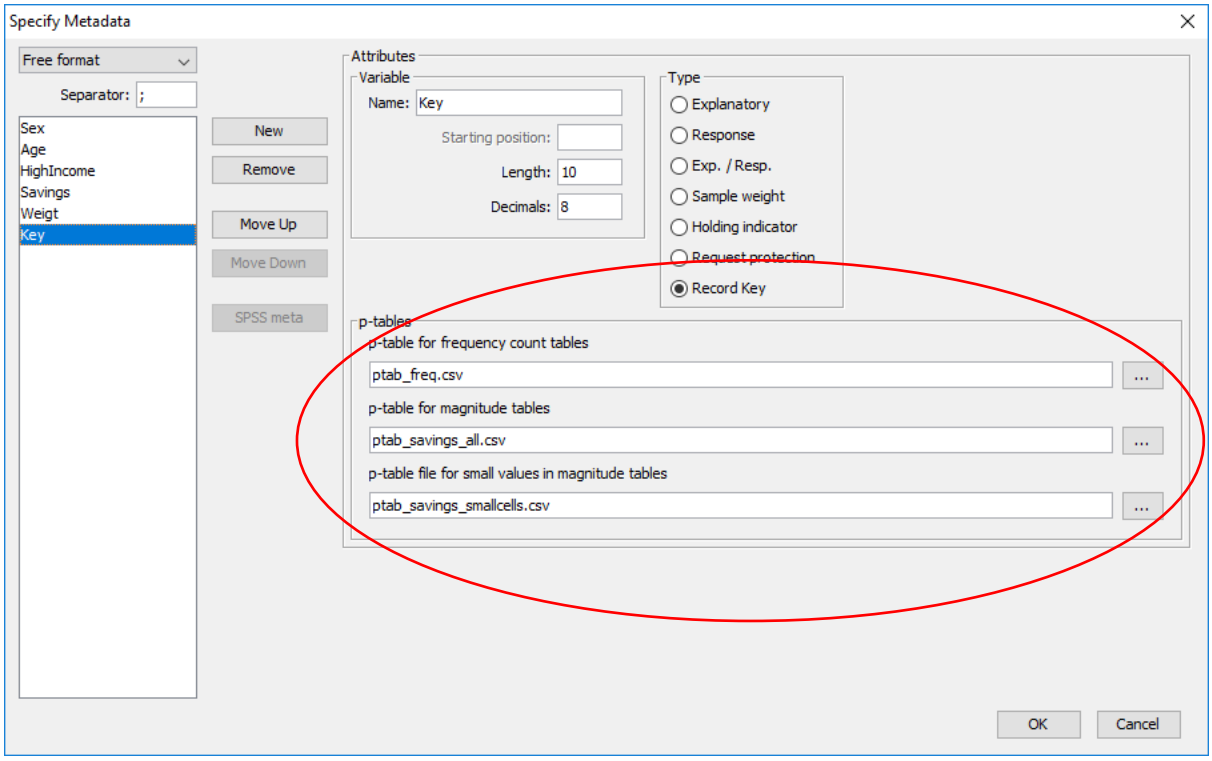
Summary of the table

Added noise	Number of cells	Relative to non-empty cells	Relative to all cells
-5	1	0,840%	0,617%
-4	3	2,521%	1,852%
-3	8	6,723%	4,938%
-2	19	15,966%	11,728%
-1	22	18,487%	13,580%
0	17	14,286%	10,494%
1	29	24,370%	17,901%
2	10	8,403%	6,173%
3	8	6,723%	4,938%
4	1	0,840%	0,617%
5	1	0,840%	0,617%
Empty	43	--	26,543%
Non-Empty	119	100%	73,457%
Total	162	--	100%

False non-zero-cells *(zero changed to non-zero)*	*0*
False zero-cells *(non-zero changed to zero)*	*8*

CKM applied to magnitude variables, a walk through

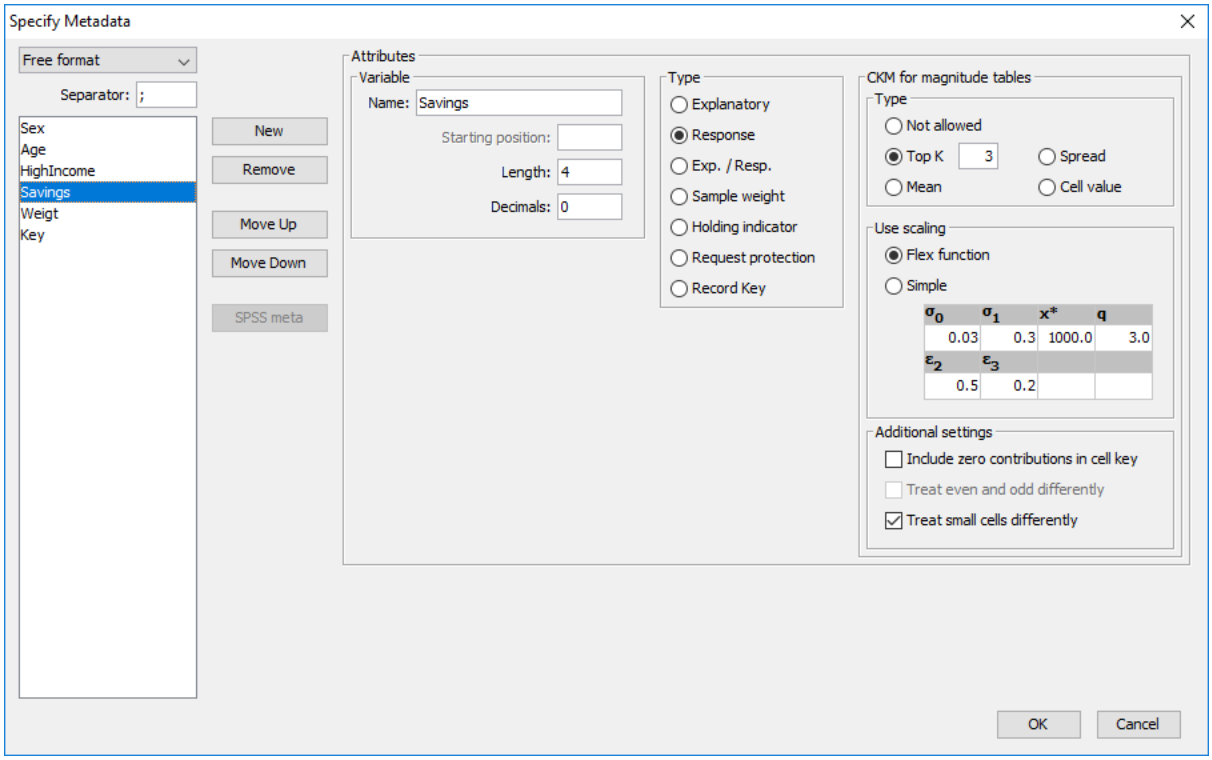
1. Open a microdata file and specify the corresponding metadata (e.g. via the `.rda`-file). In the next screenshot of 'Specify Metadata' you can see how the record-key is selected and the corresponding p-table files are selected.



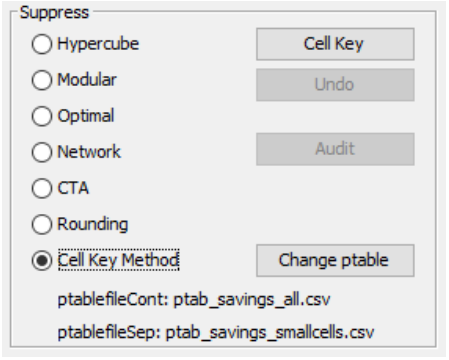
2. In the `Specify Metadata' screen, you can also inspect and specify or modfy parameters to be used for applying CKM to magnitude tables, per variable. E.g., when the `.rda` file contains

```
Savings 4
  <NUMERIC>
  <CKM> T(3)
  <INCLUDEZEROS> N
  <SEPARATION> Y
  <PARITY> N
  <SCALING> F(0.03,0.30,1000,3,0.5,0.2)
```

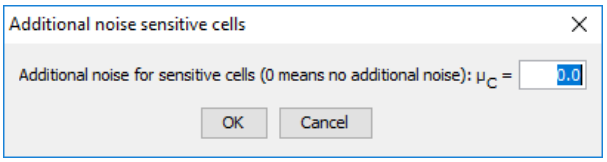
you would see



3. Specify the table to which you want to apply the cell key method. The cell key method will become available in the main screen, when you have selected a response variable that is allowed to have CKM applied to.
   You can select a sensitivity rule (e.g., the *p*%-rule) to determine sensitive cells. The noise addition using CKM can give additional protection to those sensitive cells (see the next step).
4. Select the Cell Key Method in the Suppress-panel on the main screen:

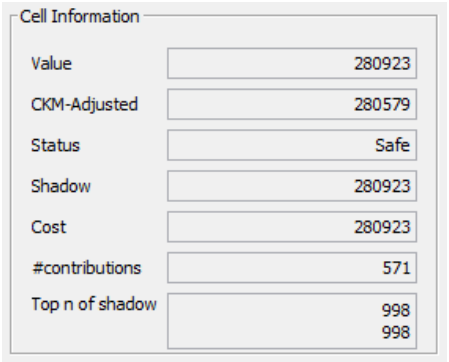


   If you want you can choose a different p-table files by pressing the 'Change ptable' button.
5. To apply the Cell Key Method, press the button labelled Cell Key. In case you have used a sensitivity rule to find sensitive cells, a pop-up will appear in which you can specify the additional amount of protection to be added to the sensitive cells:

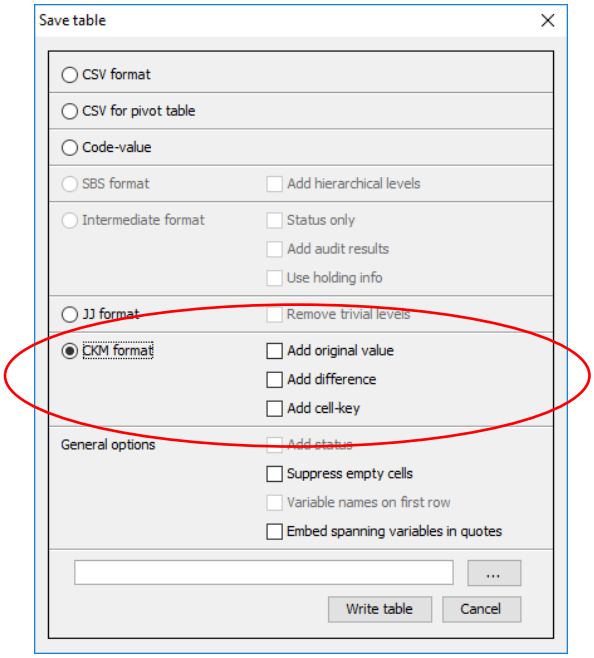


Leaving the value at 0.0 means that no additional protection is given to sensitive cells.

6. You will get the protected table without the visual feedback as in case of frequency count tables. But, obviously, you can get information on any cell by selecting that cell:



   The table summary does not yet contain information on the protected table, save for the announcement that it is protected using the cell key method.
7. You can save the protected table in the usual formats (except `Intermediate format' and `SBS format'). Additionally, you can use a new format:



   This will result in a `.tab` file where each line will contain (Italics means 'as far as selected', i.e., is optional):

   Var1; Var2; …; VarX; Cell Value*; Original Cell Value; Difference; Cell-Key*

   So e.g.,

   female;A1;464591;464726*;-135;0.12616321*

8. The report file will contain information on the applied method:

Protected with the Cell Key Method

Parameters used:

CKM Type	*T: using TopK largest with TopK = 3*
Scaling	*F: using flex function with $\sigma_0 = 0.03$, $\sigma_1 = 0.3$, $x^* = 1000.0$, $q = 3.0$, $\varepsilon_1 = 1.0$, $\varepsilon_2 = 0.5$, $\varepsilon_3 = 0.2$*
Zero contributions	*N: zero contributions are not included in recordkey*
Even/odd	*N: Even and odd number of contributors are not treated differently*
Small cells	*Y: Small cells are treated differently*

P-table used from files

p-table for general cells	*E:\[redacted]TauArgus4.2.0b5\data\cellkey\Magnitude\ptab_savings_all.csv*
p-table for small cells	*E:\[redacted]TauArgus4.2.0b5\data\cellkey\Magnitude\ptab_savings_smallcells.csv*