Instituto Complutense de Análisis Económico

Automatic regrouping of strata in the chi-square test

Juan Manuel Pérez-Salamero González

Department of Financial Economics and Actuarial Science

University of Valencia. (Spain)

Marta Regúlez-Castillo

Department of Applied Economics III

University of the Basque Country (UPV/EHU) Bilbao (Spain)

Manuel Ventura-Marco

Department of Financial Economics and Actuarial Science

University of Valencia. (Spain).

Carlos Vidal-Meliá

Department of Financial Economics and Actuarial Science, University of Valencia

and Research Institute of Economic Analysis (ICAE), Complutense University of Madrid

Abstract

Pearson´s chi-square test is widely employed in social and health science to analyze categorical data and contingency tables and to assess sample representativeness. For the test to be valid the sample size must be big enough to provide a minimum number of expected elements per category. If the researcher chooses to regroup the strata in order to solve the failure on the minimum size requirement, the existence of automatic re-grouping procedures in statistical software would be very useful, especially when tests are applied sequentially. After comprehensively reviewing the software that can carry out this test, we find that, with a few exceptions, there is no automatic regrouping of the strata to meet this requirement, although it would be very useful if this were available. This paper develops some functions for regrouping strata automatically no matter where they are located, thus enabling the test to be performed within an iterative procedure. The functions are written in Excel VBA (Visual Basic for Applications) and in Mathematica, so it would not be hard to implement them in other languages. The utility of these functions is shown by using three different datasets. Finally, the iterative use of the functions is applied to the Continuous Sample of Working Lives, a dataset that has been used in a considerable number of studies, especially on labor economics and the Spanish public pension system.

Keywords Chi-square test, statistical software, VBA, Mathematica, Continuous Sample of Working Lives

JEL Classification C46, C88, H55

Working Paper nº 1724
October, 2017

UNIVERSIDAD COMPLUTENSE MADRID

ISSN: **2341-2356**

Automatic regrouping of strata in the chi-square test

Juan Manuel Pérez-Salamero González[1], Marta Regúlez-Castillo[2], Manuel Ventura-Marco[3] and Carlos Vidal-Meliá[4]

Abstract

Pearson´s chi-square test is widely employed in social and health science to analyze categorical data and contingency tables and to assess sample representativeness. For the test to be valid the sample size must be big enough to provide a minimum number of expected elements per category. If the researcher chooses to regroup the strata in order to solve the failure on the minimum size requirement, the existence of automatic re-grouping procedures in statistical software would be very useful, especially when tests are applied sequentially. After comprehensively reviewing the software that can carry out this test, we find that, with a few exceptions, there is no automatic regrouping of the strata to meet this requirement, although it would be very useful if this were available. This paper develops some functions for regrouping strata automatically no matter where they are located, thus enabling the test to be performed within an iterative procedure. The functions are written in Excel VBA (Visual Basic for Applications) and in *Mathematica*®[5], so it would not be hard to implement them in other languages. The utility of these functions is shown by using three different datasets. Finally, the iterative use of the functions is applied to the Continuous Sample of Working Lives, a dataset that has been used in a considerable number of studies, especially on labor economics and the Spanish public pension system.

Keywords: Chi-square test, statistical software, VBA, *Mathematica*, Continuous Sample of Working Lives.

JEL: C46, C88, H55

We gratefully acknowledge financial support from Ministerio de Economía y Competitividad (Spain) and from the Basque Government via projects ECO2015-65826-P and IT 793-13 respectively. We would also like to thank Jose M. Pavía and Fernando Tusell for their comments and suggestions and Chris Pellow for his English support. Any errors are entirely due to the authors.

[1] Department of Financial Economics and Actuarial Science, University of Valencia, Avenida de los Naranjos s.n., 46022 Valencia. (Spain). (e-mail: juan.perez-salamero@uv.es).

[2] (Corresponding author) Department of Applied Economics III, University of the Basque Country (UPV/EHU). Avda. Lehendakari Aguirre 84, 48015 Bilbao (Spain). (e-mail: marta.regulez@ehu.eus)

[3] Department of Financial Economics and Actuarial Science, University of Valencia, Avenida de los Naranjos s.n., 46022 Valencia. (Spain). (e-mail: manuel.ventura@uv.es).

[4] Department of Financial Economics and Actuarial Science, University of Valencia, Avenida de los Naranjos s.n., 46022 Valencia. (Spain) and Research Institute of Economic Analysis (ICAE), Complutense University of Madrid. (e-mail: carlos.vidal@uv.es).

[5] ® Mathematica is a registered trademark of Wolfram Research Inc.

1. Introduction

Empirical studies require data samples to be representative of the target population with respect to the principal characteristics. There are many papers on the issue of selecting representative samples, including Ramsey and Hewitt (2005), Grafström and Schelin (2014), Kruskall and Mosteller (1979a, b, c, 1980), and Omair (2014).

One way of determining whether a sample is representative of a population is to use a goodness of fit test to check whether the data fits the population distribution. The goal is to test whether the sample data fits a distribution from a certain population. One procedure commonly used is Pearson´s $\mathcal{X}^2$ goodness of fit test. When the variables under study are grouped in given categories or strata in the population, the data in the sample is organized in the same way in order to apply this test. The strata are constructed so that the population is divided into major categories that are relevant to the research interest. In each category the test statistic compares the observed frequency in the sample with the expected frequency in the theoretical or known population.

Pearson's $\mathcal{X}^2$ and the likelihood ratio G^2 are arguably the two most widely used statistics in contingency table analysis (Cai *et al*. (2006)). Both can be used to test independence between categorical variables in contingency tables and to test homogeneity to determine whether frequency counts are distributed identically across different populations. These statistics may also be used to assess goodness of fit in multivariate statistics such as logistic regression (Hosmer *et al*. (1997), Hosmer and Lemeshow (2000)), log-linear modeling (Bishop *et al*. (1975), Fienberg (2006)) and Latent Class analysis (LCA) (Lazarsfeld and Henry (1968), Goodman (1974)).

These statistics have an asymptotic chi-squared distribution, given some assumptions. By and large, the validity of the test results depends on a minimum size of expected cell frequencies. As a rule of thumb, that number has been established in practice as 5. It is well-known (Cochran, 1952) that when some expected cell frequencies or probabilities are small, their reference asymptotic distribution is not suitable for assessing p-values or the size of the test. This problem arises frequently in social science, biomedical and health science and psychometrics applications (Cai *et al.* (2006), Bartholomew and Tzamourani (1999)) with sparse contingency tables (Agresti, 2002).

Delucchi (1983) reviews the research conducted after the paper by Lewis and Burke (1950) in an attempt to address the problems listed by them and to form recommendations regarding the use and misuse of the chi-square test. The various papers examined by Delucchi (1983) regarding the problem of working with too small expected frequencies recommend different minimum sizes depending on the type of test for all the strata or for a percentage of them, with fixed values or values depending on the number of categories, etc. Along the same lines, Moore (1986) establishes some criteria for the selection of the minimum size.

To solve these limitations, various alternative approaches have been proposed in the literature. One of them is to use resampling methods such as the parametric bootstrap to obtain an empirical p-value (Lin *et al*. (2015), Bartholomew and Knott (1999), Bartholomew and Tzamourani (1999); Collins *et al*. (1993)). The use of resampling methods has become increasingly popular given the power of today's computers. Cai *et al*. (2006) point out that resampling methods are not very practical computationally given that in comparing the fit of different models the resampling procedure must be repeated for each model. Moreover, Tollenaar and Mooijaart (2003) show that the validity of a bootstrap-based test depends critically on what statistic is being

bootstrapped. In particular, bootstrapping Pearson's $\mathcal{X}^2$ or the likelihood ratio G^2 does not provide immediate Type I error rate control under sparseness.

Other alternatives call for Yate´s continuity correction[6] to be used (Yates (1934)), applying exact tests such as Fisher´s exact test (Fisher (1935), Mehta and Patel (1983)) to test independence[7], or trying to estimate the Cumulative Distribution Function (CDF) of the statistics (Tsang and Cheng (2013)).

One last proposal, which has proved very popular in practice, is to pool or regroup cells to reach the desired minimum number of expected frequencies. If the test is to be conducted just once and regrouping is the option chosen (in spite of its limitations[8]), it could be carried out exogenously before the statistic is computed.

However, tests can often be used repeatedly in successive studies, or more importantly there may be techniques that use a test in an iterative process. An example of this last case would be to carry out sampling or subsampling (Pérez-Salamero *et al*. (2017)), including the goodness of fit test in mathematical programming problems. Similar examples could be found (see Marsaglia (2003)) in the analysis of random number generation processes, where tests have to be performed a number of times or in the sequential analysis of goodness of fit for different models using contingency tables. Therefore, if the researcher chooses to regroup the strata in order to solve the failure on the minimum size requirement, automatic re-grouping procedures in statistical software would be very useful, especially when tests are applied sequentially.

The paper proceeds as follows. After this introduction, Section 2 presents an extensive analysis of the software that carries out the Pearson´s $\mathcal{X}^2$ test in order to check whether there is any automatic regrouping in the strata to satisfy the desired requirement of a minimum size. We conclude that, in general, there is not. Section 3 shows the flowchart that inspired the development of the proposed functions for regrouping the strata to satisfy the desired minimum requirement, independently of whether they are in the tails or in the middle. Section 4 shows some results. Two examples are used to illustrate the utility of the functions and to analyze the behavior of the test in different software packages. Lastly, a third example shows the iterative use of the regrouping functions in a mixed integer programming framework. This is a real problem based on the Continuous Sample of Working Lives, a dataset widely used in numerous studies, especially on labor economics and the Spanish public pension system.

The paper ends with some concluding remarks, further research proposals and three appendices. The first appendix gives an outline review of selected software packages as regards whether they include Pearson´s $\mathcal{X}^2$goodness of fit test, or at least functions that enable that test to be conducted. The second shows the codes developed in *Excel VBA* (Visual Basic Applications) and *Mathematica* that make automatic regrouping and the correct application of the test $\mathcal{X}^2$ possible. The final appendix shows the mathematical approach to the real problem explained in Section 4 (Results), i.e. the selection of the larger subsample that verifies the goodness of fit $\mathcal{X}^2$ test.

[6] This correction reduces the numerical value of the test statistic, and hence weakens the power and significance level of the test, making it overly conservative (Haviland (1990), Hirji (2006), Agresti (2002), Lydersen *et. al*. (2009)).

[7] Campbell (2007) and Kroonenberg and Verbeek (2017) compare and discuss the problem of selection from these alternatives.

[8] See for example Bosgiraud (2006) or Bartholomew and Tzamourani (1999) for an excellent discussion on this issue.

2. The Chi-square Test in software

The $\mathcal{X}^2$ goodness of fit test approach can be found in any basic manual of statistical inference. It is due to the pioneer work of Pearson (1900). It is a nonparametric test which can be applied to categorical, discrete, and continuous random variables.

The statistic for the test is given by the following expression:

$$\mathcal{X}^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i} \qquad \textbf{[1.]}$$

with O_i being the observed values and E_i the expected or theoretical values. For large samples it is proved that this statistic is distributed as a $\mathcal{X}^2$ with $v = k - r - 1$ degrees of freedom, where k is the number of categories or strata, depending on how the population and the sample are organized, and "r" is the number of parameters estimated using the observed data in the sample.

The $\mathcal{X}^2$ goodness of fit test is carried out by comparing the sample value of the statistic with the corresponding critical value obtained from the $\mathcal{X}^2$ distribution with v degrees of freedom and a level α of significance. If the test statistic is not as extreme as the critical value, then the null hypothesis that the sample (observed values) has the same distribution as the population (expected values) is not rejected. The test can also be used based on the p-value obtained from the sample value of the statistic.

Table A1-1 in **Appendix 1** summarizes selected software packages that can be used for statistical purposes, to check whether Pearson´s $\mathcal{X}^2$ goodness of fit, or at least specific functions that enable it to be implemented are available in them. It also reports whether, automatic re-grouping of strata is possible if the test **[1.]** is calculated.

After comprehensively reviewing the software that can carry out this test, we conclude that there are only two programs that offer the possibility of automatic regrouping of strata when the required or desired minimum size is not reached:

a. **Matlab**, which allows users to choose the minimum size so as to regroup giving a positive integer as the value for the argument because the number zero indicates that there is no regrouping of strata in terms of the size of the expected values. The **chi2gof** function in *Matlab* regroups only the strata at the extreme end of either tail, but does not combine the interior bins.

b. **SSJ 3.2.0 Stochastic Simulation** written in *Java*. This tool allows regrouping but not in a single step. To use this facility, one must first construct an **OutcomeCategoriesChi2** object by entering the expected number of observations for each original category into the constructor. By calling up the method **regroupCategories** the program will then regroup categories in such a way that the expected number of observations in each category reaches a given threshold **minExp**. The method then counts the number of samples in each category and calls up **chi2** to get the chi-square test statistic.

There are many computer programs that have the option of filtering and/or grouping data before the test is run, but they do not offer automatic regrouping in the internal instructions for computing the test.

There is therefore consistent evidence to suggest that there are very few computer tools and statistical packages that have the possibility of automatic regrouping, not only at the extreme end of either tail but also in the interior bins. Hence, it is worth developing an automatic regrouping method that could be easily adapted to different software environments without having to perform the regrouping exogenously to the procedure each time the minimum size for the expected values is not met.

3. The automatic regrouping of strata

The automatic regrouping of categories or strata is a sequential procedure that starts with individual analysis of the size of each stratum. The second step is to regroup the categories that do not meet the minimum size requirement, if necessary, together with the adjacent ones, such that the resultants reach the desired minimum value.

It might be of interest to regroup not only the strata at the extreme ends of the tails but also those in intermediate categories. Prime examples are geographical grouping to follow some economic variables, the population at risk from certain diseases, the distribution of passengers on a track between important cities (for hours or cities with shutdown), visitor flows to shopping centers, and online submissions of tax return forms within the deadline.

In particular, the automatic strata regrouping procedure proposed analyzes their size in increasing order from the first strata to the last. The ordering is determined by the variable that is at the origin of stratification. Hence, if a category does not reach the minimum established size value, ***min***, its elements are added to the next stratum. If the next category, with the added values, reaches the minimum, ***min***, this regrouped category keeps its own elements plus the ones added, provided that the number of elements in the subsequent categories is, at least, equal to the minimum number, ***min.*** The procedure continues in the same way until the last category is reached, if applicable.

Given a specified regrouped category, if the total sum of the elements in the following categories is less than the minimum number, ***min***, those elements will be added to the given regrouped category, which becomes the last one with elements because the rest will be empty.

The flowchart in **Figure 1** shows the regrouping process on which the subsequent computation procedure is based. The main elements and the dynamics of the chart are as follows:

1. The observed and expected values needed to calculate the goodness of fit test, together with the required minimum size value for the strata, ***min***, are introduced.

2. We check whether the number of observed values in the strata, ***k***, is equal to the number of expected values, ***m***. If not, the data entry stage must be revised. If the two dimensions coincide, we continue.

3. The variable ***i***, representing the index of a specific observed and expected value, is given an initial value of 1 within the corresponding vector of values. The variable ***last*** is given an initial value of 0, and represents the indicator for the last group with a regrouped size equal to or greater than the minimum.

4. We check whether the last stratum or category has been reached by comparing the stratum index, ***i***, with the total number of strata, ***k***. If the last stratum has not yet been reached we continue with the next step. If the last stratum is reached, ***i*** = ***k***, we go to step 10.

5. The expected value in stratum $\boldsymbol{i}$, $\boldsymbol{n_i^{exp}}$, is compared with the minimum size established at the beginning, $\boldsymbol{min}$. It is worth noting that, except in the first category or stratum, the size of the expected value in a category to be compared with the minimum is that which is obtained after the loop 4-9 is performed; in other words, it might be the result of the sum of the original value for this category and previous ones which failed to reach the required minimum size.

6. If the size of the expected value in a category does not reach the minimum, this value is added to the next stratum, $\boldsymbol{n_{i+1}^{exp} = n_{i+1}^{exp} + n_i^{exp}}$. The same is done for the size of the observed value, $\boldsymbol{n_{i+1}^{obs} = n_{i+1}^{obs} + n_i^{obs}}$.

7. Once the expected and observed values of a category that does not reach the minimum size are added to the corresponding values in the next stratum, we give a value of zero to the expected and observed values of the original categories.

8. If the expected value reaches the minimum size the index that marks the last category reaching the minimum increases by 1, $\boldsymbol{last = last + 1}$, and we continue.

9. The index that marks the next category of expected values to be checked increases by 1, $\boldsymbol{i = i + 1}$, and we move back to decision node 4. The loop 4-9 is performed again, until the last group is reached, $\boldsymbol{i = k}$.

10. Once the last category of expected values is finally reached, its accumulated expected value, $\boldsymbol{n_k^{exp}}$, is compared with the minimum, $\boldsymbol{min}$.

11. If the accumulated expected value for the last category, $\boldsymbol{n_k^{exp}}$, does not reach the minimum, $\boldsymbol{min}$, the relevant value is added to the last one which did reach the minimum size, $\boldsymbol{n_{last}^{exp} = n_{last}^{exp} + n_k^{exp}}$, and the same is done with the observed value of the original last one, $\boldsymbol{n_{last}^{obs} = n_{last}^{obs} + n_k^{obs}}$. If the accumulated expected value for the last group reaches the minimum then it remains unchanged.

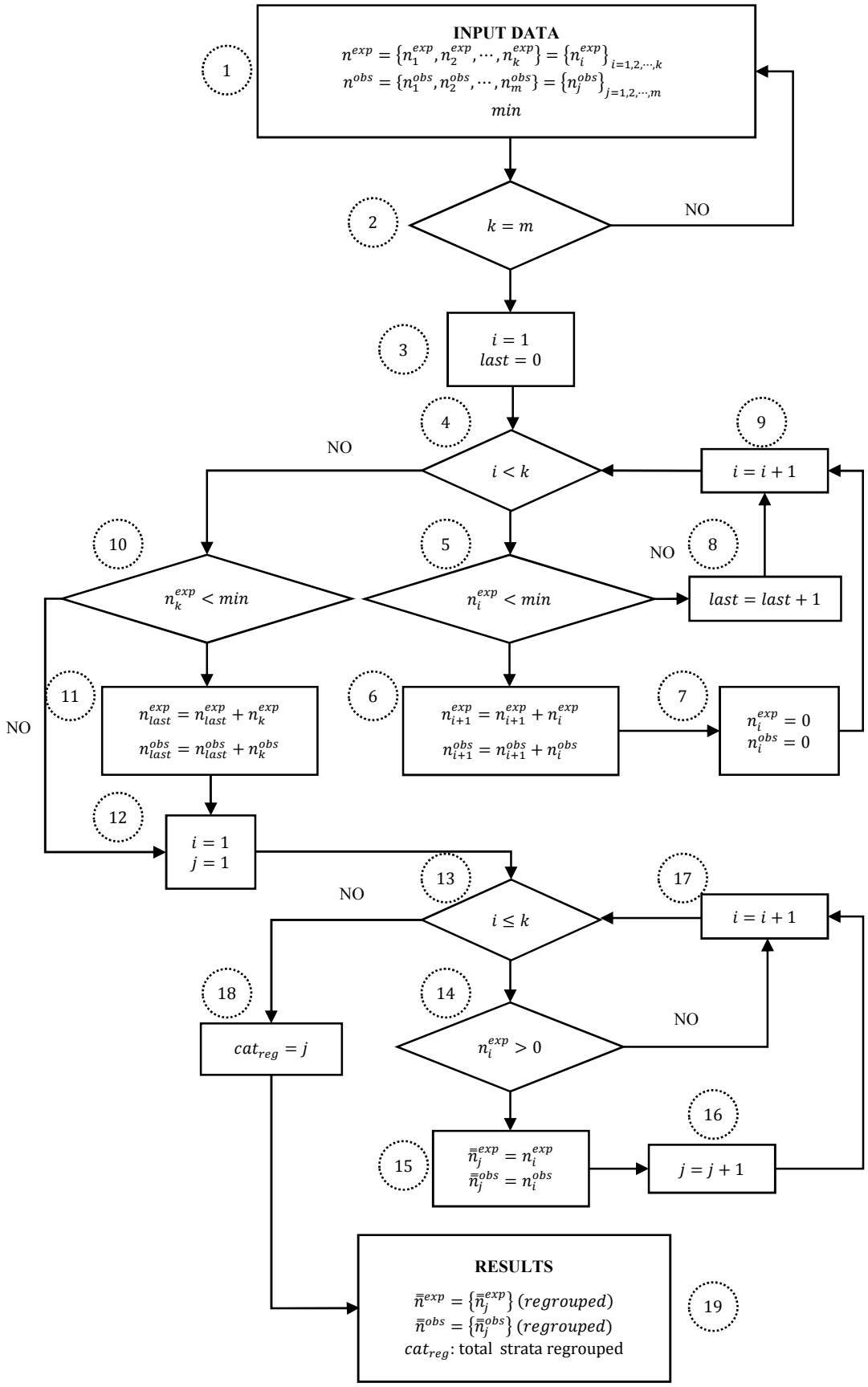
12. The indexes for the categories (original and regrouped) are initialized, $\boldsymbol{i = 1}$, $\boldsymbol{j = 1}$.

13. We start a new loop, 13-17, to compact the vector of regrouped expected and observed values obtained in the previous steps. This loop is performed for all expected values of strata, from the first to the last, $\boldsymbol{k}$, i.e. for all $\boldsymbol{i \leq k}$.

14. We check whether the accumulated expected value is greater than 0, $\boldsymbol{n_i^{exp} > 0}$, which means that, given step 7, it will be greater than the minimum. This is the updated value of the expected values for all categories. As explained above, it is obtained after adding the values of previous categories to the original value of a category, maintaining the original expected value in the case of other categories. This is done once the value of 0 has been assigned to those that do not reach the minimum and therefore do not take in the below-minimum values of other previous categories.

15. If after regrouping the accumulated expected value of the ith category is greater than 0, and therefore greater than the minimum, that value is assigned as the jth component of a new vector of regrouped expected values, $\boldsymbol{\bar{\bar{n}}_j^{exp} = n_i^{exp}}$, and the ith updated observed value is assigned to the jth component of the new vector of regrouped observed values $\boldsymbol{\bar{\bar{n}}_j^{obs} = n_i^{obs}}$.

Figure 1.- Flowchart. Automatic regrouping of strata



16. We count the regrouped number of strata compacted up to this point, adding 1 to the index variable of regrouped strata in the new vectors, $\boldsymbol{j = j + 1}$.

17. We increment the index $\boldsymbol{i}$ for the original categories of the expected values, $\boldsymbol{i = i + 1}$, up to the maximum, $\boldsymbol{k}$.

18. Once the loop 13-17 ends, the final number of regrouped categories in the new vector of expected values, $\boldsymbol{h_{reg} = j}$, is obtained.

19. The information that enables Pearson's chi square goodness of fit test to be carried out after the regrouping of strata is now available: $\bar{\bar{n}}^{exp} = \left\{\bar{\bar{n}}_j^{exp}\right\}$, $\bar{\bar{n}}^{obs} = \left\{\bar{\bar{n}}_j^{obs}\right\}$, cat_{reg}: total strata regrouped.

The procedure for the regrouping of strata or categories given a minimum size is written in *Excel Visual Basic* (*VBA*). As McCullough (2008) reports, it is well known there are quite a few shortcomings about this statistical package but he also points out, as Wilkinson (1994) and Ripley (2002) claim, that it is the most commonly used software in basic statistical calculations. This is one of the main reasons to analyze its precision, Keeling and Pavur (2011), as well as to provide functions that can be incorporated to the Microsoft Excel Function Library to help other users, as other authors have already done like Okeniyi and Okeniyi (2012) or for example improving *Excel* as a useful tool for teaching (Quintela and Fernández (2016)). In the specialized literature there is an example of using *Visual Basic* (Khan (2003)) related to Fisher's exact test (FET). This test calculates the probability value for the relationship between two dichotomous variables in a 2x2 contingency table. FET is used when a chi-square test is to be conducted but at least one of the cells has an expected frequency of five or less. FET can be used regardless of how small the expected frequency is. Khan (2003) emphasizes the potential utility of *Visual Basic* because the program is user friendly, because of its object-oriented feature and because most users are familiar with a Microsoft windows environment, especially in biomedical applications.

Furthermore, the procedure is written in *Mathematica* to illustrate that the proposed functions can be generalized to other software. As for example McCullough (2000) points out, *Mathematica* is not properly a statistical package, but it has complements to carry out the statistical analysis with more precision than others statistical packages. The functions are inspired by the work of Ross (2015) and Pérez-Salamero (2015), the latter reference being written in *VBA*. More specifically, the programming adopts functions defined by the user which yield the values for the elements needed to calculate the $\mathcal{X}^2$ test; in other words, the programming relies on the functions already available related to the test.

Tables A2-1 and **A2-2** (the latter for *Mathematica*) in **Appendix 2** show the code of the functions that yield the value of the $\mathcal{X}^2$ statistic after automatic regrouping starting from a minimum value set by the user. The length of the code corresponds more to explanatory purposes than to an effort to keep it short.

There is a difference between the functions that yield the observed and expected values in VBA and *Mathematica*. In the former we opt to define a matrix function such that the result appears in many cells because the user does not know exactly when the function will need to be used or how many regrouped categories will result. The function is written so that it selects two columns and as many rows as there were original categories such that the user can see the regrouped categories and those with zero values.

In the case of *Mathematica*, the function that returns the vectors of observed and expected values is designed to compact the categories, showing only those regrouped with values above the minimum, i.e. those with non-zero values are eliminated, as indicated in the flowchart (loop 13-17).

Table A2-3 in **Appendix 2** presents the code for the functions written in *VBA*. These functions give the number of regrouped strata in order to determine the degrees of freedom for the test. Likewise, **Table A2-4** shows the code for a matrix function in Excel which yields the output of the observed and expected values of the regrouped categories.

Finally, for the case of Mathematica **Appendix 2** incorporates the number of categories, (**Table A2-5**), the p-value for the test (**Table A2-6**), and lastly **Table A2-7** shows the relevant information set, such as the value of the $\mathcal{X}^2$ statistic, the p-value, and the regrouped strata (observed and expected).

4. Results

We use three datasets to illustrate the use of the custom functions defined in *Excel* and *Mathematica* where the regrouping of strata or categories could arise. In the first two the functions proposed in this paper are compared with some of the software tools analyzed in **Appendix 1.** Some do not regroup automatically and others, like *Matlab*, do so but only at the extreme ends of the tails. Finally, the iterative use of the regrouping functions is shown using the third data set, taking the condition that the null hypothesis for the $\mathcal{X}^2$ test is not rejected as a constraint in an optimization problem in *Excel*.

Case 1. Pearson's Ilustration V

The data labeled "Illustration V" comes from the paper by Pearson (1900). **Table 1** shows that 6 of the 17 categories considered have positive expected values of less than 5, with 4 of them being less than 1. Those strata are all located in the bins at the extreme ends.

Pearson (1900) considers that there are 17 categories, though the expected value of the last one is zero. Taking into account all the strata and with no regrouping, the sample value of the $\mathcal{X}^2$ statistic and with the 16 degrees of freedom give a p-value of 0.101. On the other hand, the functions defined in *Excel* and *Mathematica* presented in **Appendix 2** regroup them into 10 categories. The last two columns of **Table 1** show how the functions regroup the categories. Considering the 10 categories after regrouping, the sample value of the $\mathcal{X}^2$ statistic and the 9 degrees of freedom that now exist give a p-value of 0.31083538.

Excel, using the function *CHISQ.TEST*, shows the error message #¡DIV/0! because there is an expected value of 0. If group 17 is deleted the p-value is 0.073881753.

QuickCalcs in ***GraphPad*** displays the following error message "*The chi-square test is not possible when any of the expected values are zero*" if all 17 of Pearson's categories are considered. If the last category, with an expected value of 0, is deleted then there is another error message: "*The total of the observed and expected must be equal*"[9], because the program determines that the total expected values are not equal to the total observed ones.

[9] The custom functions in *Excel* VBA presented in **Appendix 2** include lines of code that verify this requirement, but they are shown as a comment to facilitate the iterative use of those functions.

Table 1. Case 1. Illustration V.				
Category	**Original**		**Regrouped**	
	Observed	**Expected**	**Observed**	**Expected**
1	0	0.18	0	0.00
2	3	0.68	0	0.00
3	7	13.48	10	14.34
4	35	45.19	35	45.19
5	101	79.36	101	79.36
6	89	96.10	89	96.10
7	94	90.90	94	90.90
8	70	71.41	70	71.41
9	46	48.25	46	48.25
10	30	28.53	30	28.53
11	15	14.94	15	14.94
12	4	6.96	10	11.34
13	5	2.88	0	0.00
14	1	1.06	0	0.00
15	0	0.34	0	0.00
16	0	0.10	0	0.00
17	0	0.00	0	0.00
Total	500	500.36	500	500.36
χ^2	11.75		10.51	
df	16		9	
p-value	0.101		0.31083538	
Source: Own work based on Pearson (1990)				

Statistica carries out the test with all 17 categories. It does not take into account any regrouping or make any requirement of minimum size for the expected values, but it warns that the sum of the observed values does not match the sum of the expected ones.

Matlab, with a default minimum size of 5, regroups the bins from the extreme ends with no warning about the difference between the sum of the observed values and the expected ones. If the parameter '**EMin**' in **chi2gof** is modified to 0 to avoid regrouping an error message is displayed, given that there is a 0 value in category 17, because we are dividing by zero. If the last category, with the expected value zero, is deleted then the same results are obtained as if there is no regrouping. This means that the data must be input in its original version, with no modifications, because it is the command that regroups given the required minimum value.

Minitab notices that there are very small values when the expected value is 0. ***R***, using the **chisq.test** function, warns users that the result might be incorrect, but does not specify that the problem is the small size of a category. It does not regroup or allow zero expected values.

Finally, ***NCSS*** does not regroup or warn about below-minimum size in any category. If it detects an expected value of 0 it does not display an error because it does not consider this kind of calculation.

Case 2. Example: "No Moore rules"

In this example, as in Case 1, the dataset does not meet the rules indicated by Moore (1986) for the minimum size required to carry out the $\mathcal{X}^2$ test. Moore establishes a general minimum size of 1, but it should be 5 in 80% of the categories. In this example the size of the expected values is below 5 in 5 of the 10 categories, and below 1 in 3 of them. Moreover, there are intermediate categories that do not satisfy the minimum size, i.e. bins 6 and 7 with values lower than 5.

Table 2. Case 2. "No Moore rules".

Category	Original		Regrouped	
	Observed	Expected	Observed	Expected
1	2	0.50	0	0.00
2	2	1.00	0	0.00
3	11	15.00	15	16.50
4	4	8.00	4	8.00
5	19	12.00	19	12.00
6	1	0.75	0	0.00
7	6	2.00	0	0.00
8	16	20.25	23	23.00
9	4	0.50	0	0.00
10	13	18.00	17	18.50
Total	78	78	78	78
$\boldsymbol{\mathcal{X}^2}$	47.51419753		6.341318591	
df	9		4	
p-value	0.0000003147		0.1750672369	

Source: Own work based on Moore (1986)

Applying the custom functions in *Excel* and *Mathematica* presented in **Appendix 2**, the data are regrouped into 5 categories. The last two columns of **Table 2** show how the categories are regrouped. Considering the 5 categories after regrouping, the sample value obtained for the $\mathcal{X}^2$ statistic and the 4 degrees of freedom give a p-value of 0.1750672369.

As in Case 1, there are problems in conducting the test in the software packages in general, because there is no automatic regrouping of the small size categories. The ways in which this issue is treated in some programs are outlined below by way of example.

GraphPad neither regroups nor performs the test, but it warns that the minimum size requirement of 5 is not satisfied: "*The chi-square calculations are only reliable when all the expected values are 5 or higher. This assumption is violated by your data, so the P value may not be very accurate*".

Matlab does not regroup the intermediate categories. It warns that the size of some of them does not reach the minimum required, so the test results may not be very reliable. Statistica neither regroups nor gives any advice about the violation of the assumptions. Lastly, ***R***, using the **chisq.test** function, warns that the results may not be very accurate but does not specify that the problem is the small size of some categories. It neither regroups nor allows for zero expected values.

Case 3. Example with the Continuous Sample of Working Lives

This case illustrates the iterative use of the regrouping functions. The $\mathcal{X}^2$ test is included as a constraint that requires that the null hypothesis not be rejected in an optimization problem written in *Excel*.

This example is taken from Pérez-Salamero *et al*. (2017). The sample data used is the Continuous Sample of Working Lives (CSWL) survey from Spain for calendar year 2013 (DGOSS, 2014). A comprehensive overview of this dataset can be found in Pérez-Salamero *et al*. (2016, 2017) and MESS (2017). The Continuous Sample of Working Lives (CSWL) is a simple random sample of around 4% of the reference population defined as individuals who have had some connection (through contributions, pensions or unemployment benefits) with the Social Security system at any time during the year of reference. It contains administrative data on working lives, which provide the basis for this sample taken from Spanish Social Security records, and comprises anonymized microdata with detailed information on individuals.

Using a post-stratification process, Pérez-Salamero *et al*. (2017) obtain from the CSWL for calendar year 2013 the data corresponding to the number of male pensioners classified as permanently disabled, organized by age in 18 categories or strata. The population distribution is known at December 31st (INSS (2014)), which means that the relative expected frequencies are also known, and hence so are the expected values. **Table 3** shows the observed values from the CSWL and the expected values from the population together with the result of regrouping.

Table 3. CSWL 2013. Permanent Disability. Males

Age Category	Original		Regrouped	
	Observed	Expected	Observed	Expected
15-19	0	0.04	0.00	0.00
20-24	29	30.04	29.00	30.08
25-29	198	195.33	198.00	195.33
30-34	606	581.48	606.00	581.48
35-39	1,201	1,203.73	1,201.00	1,203.73
40-44	2,014	1,982.02	2,014.00	1,982.02
45-49	3,106	3,050.46	3,106.00	3,050.46
50-54	4,281	4,230.30	4,281.00	4,230.30
55-59	5,710	5,706.36	5,710.00	5,706.36
60-64	7,151	7,269.83	7,151.00	7,269.83
65-69	3	58.48	3.00	58.48
70-74	6	3.28	0.00	0.00
75-79	7	4.28	13.00	7.56
80-84	14	10.88	14.00	10.88
≥ 85	17	16.48	17.00	16.48
Total	24,343	24,343	24,343	24,343
χ^2	62.76		62.66	
df	14		12	
p-value	0.0000000382		0.0000000074	

Source: Own work based on Pérez-Salamero et al. (2017)

At the bottom of **Table 3** the results of the χ^2 goodness of fit test are also shown. The null hypothesis of the test is rejected in the case of automatic regrouping[10] and in the case of no regrouping of strata with sizes lower than 5, selecting the strata with non-zero expected frequency. The ***CHISQ.TEST*** function written in *Excel* is used. Moreover, the fit of the sample to the population could be improved, since the null hypothesis is

[10] *Matlab* regroups the strata at the extreme ends but not in intermediate areas, though it warns that there are strata with numbers below the required minimum of 5.

rejected, given that the p-value is very small. If a subsample from the CSWL is selected such that its distribution does not reject the null hypothesis for a given significance level, this would provide a more representative subsample of the permanently disabled male pensioner population, by age, than the original sample.

To show the utility of the custom functions used iteratively, which enable the $\mathcal{X}^2$ test to be conducted with automatic regrouping of strata that violate the minimum size requirement, we propose an optimization problem with constraints. The aim is to find the largest subsample contained in the CSWL subject to the acceptance of the null hypothesis of equal distribution as the population. The search for the largest subsample is justified by an attempt to ensure that as few pension records as possible are missed out, so as not to overlook diversity in pensioners' working lives.

The mathematical development of the problem is shown in **Appendix 3**. It is implemented in *Excel* using the functions defined in **Appendix 2** that allow automatic regrouping**.** The problem is solved using *Solver* by *Frontline Systems*. Given its non-linearity, the method for solving the problem is *GRG Nonlinear.* Moreover, we omit the integer constraint **[6]** on the variables.

Accuracy in compliance of constraints is set to 0.0000001. We select the option "Multistart" to use the multistart method for global optimization with a population size of 100,000 and a random seed value of 100,000, using "Central" to estimate derivatives through central differencing. After solving 100,000 subproblems, a non-integer solution is reached ("*Solver* found a probability of reaching a global solution"). The solution is then rounded and it is checked that the one obtained is contained in the original sample.

Constraint **[2.]** in **Appendix 3**, related to the improvement of goodness of fit, is not satisfied by a small error of 0.0000528, the difference between the sample value of the test and the critical value at the 5% statistical significance level, and the p-value obtained is 0.0499992.

Table 4. Subsample from the CSWL 2013. Permanent disability. Males

Age Category	Original sample CSWL		Subsample			
	Observed	Observed (regrouped)	Observed rounded	Expected	Observed Rounded (regrouped)	Expected (regrouped)
15-19	0	0	0	0.02	0	0.00
20-24	29	29	13	12.98	13	13.00
25-29	198	198	85	84.41	85	84.41
30-34	606	606	252	251.27	252	251.27
35-39	1,201	1,201	521	520.15	521	520.15
40-44	2,014	2,014	858	856.46	858	856.46
45-49	3,106	3,106	1.321	1,318.14	1.321	1,318.14
50-54	4,281	4,281	1.832	1,827.97	1.832	1,827.97
55-59	5,710	5,710	2.471	2,465.80	2.471	2,465.80
60-64	7,151	7,151	3.148	3,141.39	3.148	3,141.39
65-69	3	3	3	25.27	3	25.27
70-74	6	0	2	1.42	0	0.00
75-79	7	13	0	1.85	0	0.00
80-84	14	14	5	4.70	8	7.97
≥ 85	17	17	7	7.12	7	7.12
Total	24,343	24,343	10,519	10,518.94	10,519	10,518.94
$\mathcal{X}^2$	62.76414	62.65720	21.6937921		19.6751904	
df	14	12	14		11	
p-value	0.0000000382	0.0000000074	0.0851289170		0.0499992088	

Source: Own work based on Pérez-Salamero *et al.* (2017)

The emergence of this solution, with no attention paid to the minimum size requirement for the strata, is due to the functions defined in **Appendix 2**. These functions regroup the original 15 strata into 12, with the regrouping being carried out at different times during the iterative process. This highlights the need for an automatic regrouping process because it is completely impossible to do it exogenously to the procedure in each iteration.

The results of the optimization process and the size of the strata associated with the solution obtained are presented in **Table 4**. The first two columns in **Table 4** correspond to the first and third columns of **Table 3** and we report them back in order to improve the comparison between the original sample and the sub-sample obtained. The last four columns in **Table 4** have the same structure as the ones shown in **Tables 1, 2** and **3**.

Table 4 shows that the p-value of 0.085 obtained for the $\mathcal{X}^2$ goodness of fit test in the subsample with no regrouping of strata does not able the null hypothesis to be rejected, whereas the p-value of 0.0499 obtained after regrouping is at the limit of rejection of the null. If we do not take into account the minimum size requirement for validating the test, the results could be wrong and opposite to the case of regrouping**.**

Related to this last example, Pérez-Salamero *et al*. (2017) conduct a similar analysis for the CSWL for 2010. They consider five types of pension and both genders simultaneously, and obtain the largest representative sub-sample contained in the original sample with 146 strata, reaching the last iteration and regrouping them into 115 categories to carry out the goodness of fit test. This illustrates the importance of having automatic regrouping when a large-scale iterative procedure is used.

5. Summary, conclusions and further research

In empirical studies where a Pearson´s $\mathcal{X}^2$ test is conducted, it is a common practice to regroup strata in order to attain a minimum size of expected frequencies for the test to be valid. In general, after comprehensively reviewing the software that can carry out this test, we conclude that there is no automatic regrouping of strata to meet this requirement, although it would be very useful if this were available. Having such automatic regrouping available in other packages would be of great use to researchers in many areas such as social science, biology and health science and others where this test is usually used in empirical research.

This paper proposes some functions that enable automatic regrouping to take place. This process is not only applied at the extreme ends of the tail strata, as in the case of *Matlab*, but also when intermediate categories do not meet the minimum size requirement, such as SSJ (a Java library for stochastic simulation) does.

The custom functions developed in this research have the advantage of being easier to implement than SSJ in an iterative process where the statistic must be calculated and the regrouping carried out in each iteration. This kind of process is illustrated with a real case example in the resolution of mathematical optimization problems. *Matlab* also has this advantage, but it does not allow regrouping in intermediate categories. Therefore, those functions allow to carry out Pearson's goodness of fit test with the possibility of regrouping categories, that we believe it is a quite important improvement on the current software available for basic statistical analysis, both in the case of the most used, *Excel*, as in other more precise packages like *Mathematica*.

We also believe that it could be very useful to make the automatic regrouping of categories or strata available in the iterative use of the test statistics used in Big Data and Data Mining (Larose (2014)), for example at the instance selection and association analysis stages, etc.

Finally, based on this paper, one possible direction for future research would be to adapt the code of the proposed functions to other languages and optimization environments such as AMPL, GAMS, LINGO, etc., in order to integrate them into the numerical resolution of problems of this kind. It would also be interesting to automate regrouping based on other, more general criteria such as sample size or the number of categories.

6. References

Agresti, A. (2002). Categorical Data Analysis (2^{nd} edn). Wiley: Hoboken, New Jersey.

Bartholomew, D. J.; Knott, M. (1999). Latent variable models and factor analysis (2nd ed.). London: Arnold.

Bartholomew, D. J.; Tzamourani, P. (1999). The goodness of fit of latent trait models in attitude measurement. *Sociological Methods and Research* 27, 525-546. DOI: 10.1177/0049124199027004003

Bishop, Y. M. M.; Fienberg, S. E.; Holland, P. W. (1975). Discrete multivariate analysis: theory and practice. MIT Press, Cambridge.

Bosgiraud, J. (2006). Sur le regroupement des classes dans le test du Khi-2. *Revue Romaine de Mathématiques Pures et Appliquées*, 51 (2), 167-172.

Cai, L.; Maydeu-Olivares, A.; Coffman, D.L.; Thissen, D. (2006). Limited-information goodness-of-fit testing of item response theory models for sparse 2^p tables. *British Journal of Mathematical and Statistical Psychology*, 59 (1), 173-194. DOI: 10.1348/000711005X66419.

Campbell, I. (2007). Chi-squared and Fisher–Irwin tests of two-by-two tables with small sample recommendations, *Statistics in Medicine* 26 (19), 3661-3675. DOI: 10.1002/sim.2832.

Cochran, W .G. (1952). The $\chi 2$ test of goodness of fit, *The Annals of Mathematical Statistics*. 23 (3), 315-345. http://www.jstor.org/stable/2236678

Collins, L. M.; Fidler, P. L.; Wugalter, S. E.; Long, J. (1993). Goodness-of-fit testing for latent class models. *Multivariate Behavioral Research*, 28 (3), 375-389. DOI: 10.1207/s15327906mbr2803_4.

Delucchi, K. L. (1983). The Use and Misuse of Chi-Square: Lewis and Burke Revisited. *Psychological Bulletin*, 94 (1), 166-176. DOI: 10.1037/0033-2909.94.

Dirección General de Ordenación. Secretaría de Estado de la Seguridad Social (DGOSS). (2014), Muestra Continua de Vidas Laborales 2013. Madrid: Ministerio de Trabajo e Inmigración. España.

Fienberg, S. E. (2006). Log-linear Models in Contingency Tables. Encyclopedia of Statistical Sciences. 7. Wiley, New York.

Fisher R. A. (1935). The logic of inductive inference. *Journal of the Royal Statistical Society,* 98 (1), 39-54. DOI: 10.2307/2342435.

Goodman, L. A. (1974). Exploratory Latent Structures analysis Using Both Identifiable and Unidentifiable Models. *Biometrika*, 61 (2),215-231. DOI: 10.2307/2334349.

Grafström, A.; Schelin, L. (2014). How to Select Representative Samples. *Scandinavian Journal of Statistics*, 41 (2), 277-290. DOI: 10.1111/sjos.12016.

Haviland, M. G. (1990). Yates's correction for continuity and the analysis of 2×2 contingency-tables. *Statistics in Medicine* 9 (4), 363-367. DOI: 10.1002/sim.4780090403.

Hirji, K. F. (2006). Exact Analysis of Discrete Data. Chapman & Hall: Boca Raton.

Hosmer, D. W.; Hosmer, T.; Le Cessie, S.; Lemeshow, S. (1997). A comparison of goodness-of-fit tests for the logistic regression model. *Statistics in medicine*, 16 (9), 965-980. DOI: 10.1002/(SICI)1097-0258(19970515)16:9<965::AID-SIM509>3.0.CO;2-O.

Hosmer D.W.; Lemeshow, S. (2000). Applied logistic regression, 2nd edn. Wiley, New York.

INSS. (2014). Informe Estadístico 2013. Madrid: INSS. Secretaría de Estado de la Seguridad Social. Ministerio de Empleo y Seguridad Social.

Keeling, K. B.; Pavur, R. J. (2011). Statistical Accuracy of Spreadsheet Software. *The American Statistician*, 65 (4), 265-273. DOI: 10.1198/tas.2011.09076

Khan, H.A. (2003). A Visual Basic Software for Computing Fisher's Exact Probability. *Journal of Statistical Software*, 8 (21), 1-7. DOI: 10.18637/jss.v008.i21.

Kroonenberg, P.M.; Verbeek, A. (2017). The Tale of Cochran's Rule: My Contingency Table has so Many Expected Values Smaller than 5, What Am I to Do? *The American Statistician*. DOI: 10.1080/00031305.2017.1286260.

Kruskall, W.; Mosteller, F. (1979a). Representative Sampling, I. *International Statistical Review/ Revue Internationale de Statistique*, 47 (1), 13-24. DOI: 10.2307/1403202.

Kruskall, W.; Mosteller, F. (1979b). Representative Sampling, II: Scientific Literature, Excluding Statistics. *International Statistical Review / Revue Internationale de Statistique*, 47 (2), 111-127. DOI: 10.2307/1402564.

Kruskall, W.; Mosteller, F. (1979c). Representative Sampling, III: The current Statistical Literature. *International Statistical Review / Revue Internationale de Statistique*, 47 (3), 245-265. DOI: 10.2307/1402647.

Kruskall, W.; Mosteller, F. (1980). Representative Sampling, IV: The History of the Concept in Statistics, 1895-1939. *International Statistical Review / Revue Internationale de Statistique*, 48 (2), 169-195. DOI : 10.2307/1403151.

Larose, D. T.; Larose C. D. (2014). Discovering knowledge in data: an introduction to data mining. John Wiley y Sons. DOI : 10.1002/9781118874059

Lazarsfeld, P.F.; Henry, N.W. (1968). Latent Structure Analysis, Houghton Mifflin, Boston.

Lewis, D. & Burke, C. J. (1949). The Use and Misuse of Chi-Square. *Psychological Bulletin*, 46 (6), 433-489. DOI : 10.1037/h0059088.

Lin, J. J.; Chang, C. H. & Pal, N. (2015). A Revisit to Contingency Table and Tests of Independence: Bootstrap is Preferred to Chi-Square Approximations as Well as Fisher's exact test. *Journal of Biopharmaceutical Statistics*, 25 (3), 438-458. DOI: 10.1080/10543406.2014.920851.

Lydersen, S.; Fagerland, M.W. & Laake, P (2009). Tutorial in Biostatistics. Recommended tests for association in 2×2 tables. *Statistics in Medicine,* 28 (7), 1159–1175. DOI: 10.1002/sim.3531. DOI: 10.1002/sim.3531.

Marsaglia, George (2003) "Random Number Generators," Journal of Modern Applied Statistical Methods, 2 (1) , 2-13. DOI: 10.22237/jmasm/1051747320. DOI: 10.22237/jmasm/1051747320.

McCullough, B. D. (2000). The Accuracy of Mathematica 4 as a Statistical Package. *Computational Statistics*, 15 (2), 279-299. DOI:10.1007/PL00022713

McCullough, B. D. (2008). Special section on Microsoft Excel 2007. *Computational Statistics and Data Analysis*, 52 (10), 4568-4569. DOI:10.1016/j.csda.2008.03.009.

Mehta C.R.; Patel N.R. (1983). A Network Algorithm for Performing Fisher's Exact Test in r x c Contingency Tables. *Journal of the American Statistical Association*. 78 (382), 427-434. DOI: 10.2307/2288652

Ministerio de Empleo y Seguridad Social, Secretaría de Estado de Seguridad Social (MESS). (2017), "La Muestra Continua de Vidas Laborales. Guía del contenido". Estadísticas, Presupuestos y Estudios. Estadísticas. [Última consulta: 8-4-2017]: http://www.seg-social.es/prdi00/groups/public/documents/binario/190489.pdf.

Moore, D. (1986). "Test of Chi-Squared Type" in D'Agostino, R. & Stephens, M., eds. *Goodness of Fit Techniques*, Marcel-Decker, New York, 63-95.

Okeniyi, J. O. ; Okeniyi, E. T. (2012). Implementation of Kolmogorov–Smirnov P-value computation in Visual Basic: implication for Microsoft Excel library function. *Journal of Statistical Computation and Simulation*, 82 (12), 1727-1741. DOI:10.1080/00949655.2011.593035.

Omair, A. (2014). Sample size estimation and sampling techniques for selecting a representative sample. *Journal of Health Specialties*, 2 (4), 142-147. DOI: 10.4103/1658-600X.142783.

Pearson, K. (1900). On the Criterion that a given System of Deviations from the Probable in the Case of a Correlated System of Variables is such that it can be reasonably supposed to have arisen from Random Sampling. *Philosophical Magazine*, 50 (5), 157-175. DOI: 10.1080/14786440009463897.

Pérez-Salamero González, J.M., Regúlez-Castillo, M. & Vidal-Meliá, C. (2017), The Continuous Sample of Working Lives: improving its representativeness. *SERIEs. Journal of the Spanish Economic Association*, 8 (1), 43-95. DOI: 10.1007/s13209-017-0154-0.

Pérez-Salamero González, J.M., Regúlez-Castillo, M. & Vidal-Meliá, C. (2016), Análisis de la representatividad de la MCVL: el caso de las prestaciones del sistema público de pensiones. *Hacienda Pública Española (Review of Public Economics)*, 217-(2/2016): 67-130

Pérez-Salamero González, J. M. (2015). La Muestra Continua de Vidas Laborales (MCVL) como fuente generadora de datos para el estudio del sistema de pensiones. Tesis Doctoral. Universitat de València.

Quintela-del-Río, A. Fernández, M. F. (2016). Excel Templates: A Helpful Tool for Teaching Statistics. *The American Statistician*. DOI: 10.1080/00031305.2016.1186115

Ramsey, C. A.; Hewitt, A. D. (2005). A Methodology for Assessing Sample Representativeness. *Environmental Forensics*, 6, 71-75. DOI: 10.1080/15275920590913877

Ripley, Brian D. (2002). Statistical methods need software: A view of statistical computing. Opening lecture RSS, Plymouth.

Ross, A. (2015). Probability or statistics-Performing a chi-square goodness of fit test-Mathematical Stack Exchange. (Retrieved: 30/4/2016) url: http://mathematica.stackexchange.com/questions/5579/permorfing-a-chi-square-goodness-of-fit-test/5590#5590

Tsang, W.W.; Cheng, K.H. (2006). The Chi-square test when the expected frequencies are less than 5, COMPSTAT 2006 - Proceedings in Computational Statistics, edited by A. Rizzi and M. Vichi, Physica Verlag (Springer), pp. 1583-1589.

Tollenaar, N.; Mooijaart, A. (2003). Type I errors and power of the parametric bootstrap goodness-of-fit test: Full and limited information. *British Journal of Mathematical and Statistical Psychology*, 56 (2), 271-288. DOI: 10.1348/000711003770480048.

Wilkinson, L. (1994). Practical guidelines for testing statistical software. In Dirschedl, P. and Ostermann, R. (Eds.), Computational Statistics. Heidelberg: Physica-Verlag

Yates, F. (1934). Contingency tables involving small numbers and the χ^2 test. *Journal of the Royal Statistical Society*, Suppl.1, 217-235. DOI: 10.2307/2983604.

Appendix 1: Pearson's goodness of fit test in the software analyzed

Table A1-1. Pearson's goodness of fit test in the software.

Software	Chi-square Test	Regroup and/or automatic regrouping
EXCEL *MicroSoft Corporation*	*CHISQ.TEST* Returns the p-value obtained from Pearson's chi-square statistic **[1.]**. If any expected value is zero the error message "#¡DIV/0! division by zero#" is displayed. https://support.office.com/en-us/article/CHISQ-TEST-function-2e8a7861-b14a-4985-aa93-fb88de3f260f	NO
GraphPad *GraphPad Software, Inc*	QuickCalcs. On-line. It computes the statistic keeping the original observed and expected values, but only warns about the violation of the requirement of a minimum size of 5. If any of the expected values is zero the test is not carried out. http://graphpad.com/quickcalcs/chisquared1/	NO
JMP *SAS Institute Inc. Cary*	If any expected value is zero, the chi-square statistic is computed without taking this into account. It reports the error but does not compute the p-value. If the expected values or expected frequencies do not add up to 1 it allows them to be rescaled.	NO
Mathematica *Wolfram Research, Inc.*	PearsonChiSquareTest is a function that computes Pearson's chi-square statistic but based on a method due to D'Agostino and Stephens. In this method the histograms of the observed and expected values are compared, so it does not calculate the statistic in the same way as Pearson. (Ross, 2015). http://mathematica.stackexchange.com/questions/5579/performing-a-chi-square-goodness-of-fit-test	NO
Matlab *The MathWorks Inc.*	chi2gof: It computes the goodness of fit statistic. It regroups the strata at the extreme ends of the tails but not the intermediate ones. It allows the minimum size requirement to be set by using the EMin option. It returns the statistic value, the regrouped strata values, and any other information required about the test. https://es.mathworks.com/help/stats/chi2gof.html	YES but only at the extreme ends of the tails.
Minitab *Minitab Inc.*	Mac: Statistics > Tables > Chi-Square Goodness-of-Fit; PC: STATISTICS > Chi-Square Goodness-of-Fit. It warns that results may not be accurate when the strata of expected values have sizes lower than 5 and 1. It also gives information about the percentage of strata that do not satisfy the requirement. The expected frequencies must be entered in place of the expected values. http://support.minitab.com/en-us/minitab-express/1/help-and-how-to/basic-statistics/tables/chi-square-goodness-of-fit-test/before-you-start/data-considerations/	NO
NCSS *NCSS, LLC*	It does not inform about the minimum size reached by some categories and does not display any error message if it detects a zero expected value, given that it does not consider such groups in carrying out its calculations.	NO
PH-Stat *Pearson Education, Inc.*	Add-in in Excel for statistical analysis based on functions written in Excel, so it has the same characteristics. In the chi-square goodness of fit tests it displays an error message about the violation of the assumption on the minimum expected frequency if it does not reach a minimum of 1 or 5 (chi-square test about difference between proportions), depending on the cases.	NO

Table A1-1. Pearson's goodness of fit test in the software.

Software	Chi-square Test	Regroup and/or automatic regrouping
PSPPIRE *Free Software Foundation, Inc.*	Free software similar to SPSS, but it does not display any minimum size requirement error message as SPSS does.	NO
R *The R Foundation for Statistical Computer*	**chisq.test**: If there are strata with expected values lower than 5, it reports that the results are not correct: "Chi-squared approximation may be incorrect". It there is a zero expected value, it does not compute the statistic because of division by zero. https://stat.ethz.ch/R-manual/R-devel/library/stats/html/chisq.test.html The statistic given by gofTest of the ENVStats is based on chisq.tets. http://finzi.psych.upenn.edu/R/library/EnvStats/html/gofTest.html	NO
SAS/STAT *SAS Institute Inc.*	The statement TABLES given in the procedure PROC FREQ does not allow for zero expected values in the option TESTF. It calculates the statistic showing the percentage of bins with expected values lower than 5 and it warns that the chi-square test results are not valid. https://support.sas.com/documentation/cdl/en/statug/63347/HTML/default/viewer.htm#statug_freq_a0000000658.htm	NO
S-PLUS[11] *TIBCO Software Inc.*	Chisq.test works as in R, given that both use the same language and many of the functions of S-PLUS. The function chisq.gof in S-PLUS computes the goodness of fit test, but for theoretical expected values given by the usual statistical distributions.	NO
SPSS *IBM Corporation y otros*	If any expected frequency is zero the test procedure stops and reports this by telling the user that the expected values in each category must be at least 1 and no more than 20% of the categories may be lower than 5. If any observed value is zero but the corresponding expected value is not, the test is not computed because it considers that there are fewer observed categories than expected. It does not allow an expected value for frequencies of less than 0.0001	NO
STATA *Estima Inc.*	To compute the statistic it deletes the zero expected values. To use the frequency tables (tabi, tab2, tabulate), it demands integer values, because it requires counts (absolute frequency), not relative frequencies.	NO
Statistica *Dell Inc.*	It reports a problem if the total sum of observed values does not coincide with that of the expected values, but it does not warn about the problem of the violation of the minimum size requirement.	NO
Stochastic Simulation in Java. SSJ 3.2.0 *L'Ecuyer et al. Universty of Montreal*	It conducts Pearson's goodness of fit test with the possibility of regrouping the strata given a minimum size required value to be set by the user. This is done in two steps using the observed and expected values. It combines the functions OutcomeCategoriesChi2 and regroupCategories. http://umontreal-simul.github.io/ssj/docs/master/classumontreal_1_1ssj_1_1gof_1_1GofStat_1_1OutcomeCategoriesChi2.html#details	YES

[11] S-Plus is has recently been integrated into the TIBCO Spotfire analytics platform with built-in data wrangling, which delivers AI-driven, visual, geo, and streaming analytics.

Table A1-1. Pearson's goodness of fit test in the software.

Software	Chi-square Test	Regroup and/or automatic regrouping
ViSta *Forrest W. Young.* **XLISP-STAT** *Luke Tierney* **XLISP version** *David Betz*	LispStat is no longer developed, because its creator is now a member of the R core team of programmers. http://archives.math.utk.edu/software/msdos/statistics/xlisp-stat/ There is software based on XLisp-Stat, such as ViSta, that warns about the existence of expected frequencies lower than 6 and that the chi-square test will not be valid. http://www.uv.es/visualstats/Book/DownloadBook.htm	NO
XLStatistics *Rodney Carr*	Add-in in Excel for statistical analysis. It works with the functions given in Excel, so it has the same problems: it does not warn about the violation of the minimum size requirement and it does not allow for zero expected values. www.deakin.edu.au/~rodneyc/xlstatistics	NO

Appendix 2: VBA in Excel[12] and Mathematica[13]. Codes[14].

Table A2-1. VBA in Excel code. Chi-square statistic with regrouped strata.

Table A2-2. Mathematica code. Chi-square statistic with regrouped strata.

Table A2-3. VBA in Excel code. Number of regrouped strata.

Table A2-4. VBA in Excel code. Observed and expected values.

Table A2-5. Mathematica code. Number of regrouped strata.

Table A2-6. Mathematica code. P-value.

Table A2-7. Mathematica code. Summary of Chi-square Test results.

Appendix 3: Case 3. Selection of the largest sub-sample that verifies the χ^2 goodness of fit test: Mathematical procedure.

$$\underset{n_i^{SUB}}{\text{Max}}\left\{n^{SUB} = \sum_{i=1}^{k} n_i^{SUB}\right\} \quad \textbf{[2.]}$$

subject to:

$$\chi^2(n_1^{SUB}, \cdots, n_k^{SUB}) = \sum_{j=1}^{cat_{reg}} \frac{(\bar{n}_j^{SUB} - \bar{n}_j^{exp})^2}{\bar{n}_j^{exp}} \leq \chi^2_{(\alpha,\ r)} \quad \textbf{[3.]}$$

$$n_i^{exp} = \frac{N_i}{N} \cdot n^{SUB} = \frac{N_i}{N} \cdot \sum_{i=1}^{k} n_i^{SUB} \quad \textbf{[4.]}$$

$$0 \leq n_i^{SUB} \leq N_i \quad \textbf{[5.]}$$

$$0 \leq n_i^{SUB} \leq n_i^{RS} \quad \textbf{[6.]}$$

$$n_i^{SUB} \in Z;\ \forall i = 1, 2, \cdots, k \quad \textbf{[7.]}$$

with

n^{SUB}: Subsample size.

n_i^{SUB}: Size of category *i* from the subsample (observed values).

k: Number of strata on the variable of interest from which the stratification is made.

$\chi^2(n_1^{SUB}, \cdots, n_k^{SUB})$: Chi-square goodness of fit test statistic. Its value depends on the size of the regrouped strata.

[12] Microsoft Visual Basic for Applications 7.1. © Microsoft Corporation 2012.
[13] Wolfram Mathematica 11.0.0.0. Microsoft Windows (64-bit).
[14] Codes are available upon request to the authors.

n_i^{exp}: Expected value size of category *i* from the subsample. It depends on the relative frequency of the population and the size of the subsample.

$\chi^2_{(\alpha,\ r)}$: Critical value from the chi-square distribution with r degrees of freedom and a given statistical significance level α fixed at 5%.

N_i: Size of stratum *i* from male pensioners classified as permanently disabled, given by INSS (2014).

$\bar{\bar{n}}_j^{exp}$: Expected size of the regrouped category j from the sub-sample. It depends on the relative frequency from the population and from the size of the subsample.

$\bar{\bar{n}}_j^{SUB}$: Proposed observed size for the regrouped stratum j from the subsample.

N: Total number of male pensioners classified as permanently disabled in the population of pensioners given by INSS (2014).

cat_{reg}: Number of regrouped strata.

$r = cat_{reg} - 1$: Degrees of freedom for the test, equal to the number of regrouped strata minus 1, given that in this case there are no parameters to be estimated because the population distribution is known.

n_i^{RS}: Size of category *i* from the post-stratification of the CSWL (Random Sample).

Z: Set of integer numbers.

Constraint **[3.]** is intended to achieve a better fit of the extracted subsample than the original (CSWL), given that it provides a value for the goodness of fit statistic that does not reject the null hypothesis. Using the functions shown in Appendix 2 the statistic and the degrees of freedom are calculated from the values of the automatically regrouped categories.

Rule **[4.]** establishes that the regrouped expected value of each category or stratum, in each iteration, automatically adapts to the new size that the subsample can take.

Constraint **[5.]** is set to prevent the outliers found in the CSWL. Given the procedure for obtaining the CSWL, and given that it comes from administrative records, the processing date of the CSWL is later than the one on which the Spanish Social Security Institute drawn up its statistics (INSS (2014)). Therefore, there might be strata in the CSWL with pensioners who do not belong to the population because their benefits have been awarded retroactively. This constraint can be ignored if the sample is really contained in the population.

Constraint **[6.]** implies that the subsample must be contained in the CSWL and **[7.]** requires that the number of pensioners to be included in each stratum of the subsample be a natural number (non-negative integer).