SORT 43 (1) January-June 2019, 113-142 DOI: 10.2436/20.8080.02.83

# Automatic regrouping of strata in the goodness-of-fit chi-square test

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

**Abstract**

Pearson's chi-square test is widely employed in social and health sciences to analyse categorical data and contingency tables. For the test to be valid, the sample size must be large enough to provide a minimum number of expected elements per category. This paper develops functions for regrouping strata automatically, thus enabling the goodness-of-fit test to be performed within an iterative procedure. The usefulness and performance of these functions is illustrated by means of a simulation study and the application to 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 labour economics and the Spanish public pension system.

*MSC:* 62G10, 62P25.

*Keywords:* Goodness-of-fit chi-square test, statistical software, Visual Basic for Applications, Mathematica, Continuous Sample of Working Lives.

## 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), Grafstörm and Schelin (2014), Kruskal and Mosteller (1979a), Kruskal and Mosteller (1979b), Kruskal and Mosteller (1979c), Kruskal and Mosteller (1980), Omair (2014). One way of determin-

[1] (Corresponding author) Department of Applied Economics III (Econometrics and Statistics), Faculty of Economics and Business, University of the Basque Country UPV/EHU. Avda. Lehendakari Aguirre 83, 48015 Bilbao. (Spain). vicente.nunezanton@ehu.eus, marta.regulez@ehu.eus.

[2] Department of Financial Economics and Actuarial Science. Faculty of Economics. University of Valencia. Avenida de los Naranjos s.n., 46022 Valencia. (Spain). juan.perez-salamero@uv.es, manuel.ventura@uv.es

[3] Department of Financial Economics and Actuarial Science. Faculty of Economics. University of Valencia. Avenida de los Naranjos s.n., 46022 Valencia. (Spain) and research affiliation with the Instituto Complutense de Análisis Económico (ICAE), Complutense University of Madrid (Spain), and the Centre of Excellence in Population Ageing Research (CEPAR), UNSW (Australia). carlos.vidal@uv.es

Received: May 2018

Accepted: January 2019

ing 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 $\chi^2$ goodness-of-fit test. When the variables under study are grouped in given categories or strata in the population, the data in the sample are 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 $\chi^2$ and the likelihood ratio test statistic $G^2$ are arguably the two most widely used statistics in contingency table analysis (see 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 in logistic regression (Hosmer et al. 1997, Hosmer and Lemeshow 2000), log-linear modelling (Bishop, Fienberg and Holland, 1975, Fienberg 2006) and Latent Class Analysis (LCA) (Lazarsfeld and Henry 1968, Goodman 1974). Under some conditions, these statistics have an asymptotic chi-square distribution, where the validity of the test results depends on a minimum size of expected cell frequencies. As a rule of thumb, that number is 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 sciences, biomedical and health sciences and psychometrics applications (Cai et al. 2006, Bartholomew and Tzamourani 1999) with sparse contingency tables (Agresti 2002).

Delucchi (1983) reviewed the research conducted after the paper by Lewis and Burke (1949) 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 excessively 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) and Wickens (1989) established some criteria for the selection of the minimum size. García Pérez and Nuñez-Antón (2009) found, via simulation, that Pearson's $\chi^2$ was sufficiently accurate and only showed minor misbehaviour when table density was less than two observations per cell for testing independence or homogeneity in two-way contingency tables. 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, Chang and Pal, 2015, Bartholomew, Knott and Moustaki, 2011, 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) pointed out that resampling methods are not very practical from a compu-

tational perspective given that in comparing the fit of different models the resampling procedure must be repeated for each model. Moreover, Tollenaar and Mooijaart (2003) showed that the validity of a bootstrap-based test depends critically on what statistic is being bootstrapped. In particular, bootstrapping Pearson's $\chi^2$ or the likelihood ratio test statistic $G^2$ does not provide immediate Type I error rate control under sparseness.

Other alternatives call for Yate's continuity correction[1] to be used (Yates 1934), applying exact tests such as Fisher's exact test (Fisher 1935, Mehta and Patel 1983) to test independence[2], or trying to estimate the cumulative distribution function (CDF) of the statistics (Tsang and Cheng 2006). 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[3]), 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 the latter would be to carry out sampling or subsampling (Pérez-Salamero González, Regúlez-Castillo and Vidal-Meliá, 2017), including the goodness-of-fit test in mathematical programming problems. Similar examples could be found (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 researchers choose to regroup the strata in order to solve the failure on the minimum size requirement in the goodness-of-fit chi-square test, automatic re-grouping procedures in statistical software would be very useful, especially when tests are applied sequentially.

The paper is organized as follows. Section 2 presents an example to motivate the problem to be solved, and extensively analyse the software that carries out the Pearson's $\chi^2$ goodness-of-fit 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 simulation results to illustrate the performance of the procedure in terms of nominal significance levels under different settings. Section 5 presents three more examples: one to illustrate the utility of the functions and to analyse the behaviour of the test in different software packages, a second to illustrate the use of the regrouping functions

1. 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, Fagerland and Laake, 2009).

2. Campbell (2007) and Kroonenberg and Verbeek (2018) compare and discuss the problem of selection from these alternatives.

3. See for example Bosgiraud (2006) and Bartholomew and Tzamourani (1999) for an excellent discussion on this issue.

when it is necessary to estimate parameters of the distribution and finally, an example that 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 (CSWL), a dataset widely used in numerous studies, especially on labour economics and the Spanish public pension system. The paper ends with some concluding remarks and further research proposals. In addition, we provide three appendices in the supplementary material. The first appendix provides a summarized review of selected software packages as regards whether they include Pearson's $\chi^2$ goodness-of-fit test, or at least functions that enable that test to be conducted. The second includes the mathematical approach to the real problem explained in Section 5, i.e. the selection of the larger subsample that verifies the goodness-of-fit $\chi^2$ test. The authors can be contacted to supply the codes developed in Microsoft Excel 2016 and Microsoft Excel VBA (Visual Basic for Applications 7.1) and Mathematica[4] that make automatic regrouping and the correct application of the $\chi^2$ test possible.

## 2. Illustration of the problem and software review

The $\chi^2$ goodness-of-fit test approach can be found in any basic manual of statistical inference. It is due to the pioneering 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:

$$\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}, \tag{1}$$

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 under the null hypothesis as a $\chi^2$ 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 $\chi^2$ goodness-of-fit test is carried out by comparing the sample value of the statistic with the corresponding critical value obtained from the $\chi^2$ distribution with $v$ degrees of freedom and a level $\alpha$ of significance. If the test statistic is less than 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.

To illustrate the problem that we seek to address with our procedure, we propose the following example that we call "No Moore rules." In this example, the dataset does not

4. Mathematica is a registered trademark of Wolfram Research Inc. version 11.

meet the rules indicated by Moore (1986) for the minimum size required to carry out the $\chi^2$ goodness-of-fit test. Moore established a general minimum size of 1, but it should be 5 in 80% of the categories. As shown in Table 1, 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 requirement, i.e. bins 6 and 7 with values lower than 5. The population distribution used is a multinomial with probabilities as shown in the second column of Table 1.

***Table 1:*** *Example featuring "No Moore rules" conditions. Values for the goodness-of-fit $\chi^2$ test statistic, degrees of freedom (df) and p-values are also reported.*

| Category | Pop. prob. | Original | | Regrouped | |
|---|---|---|---|---|---|
| | | Obs. | Exp. | Obs. | Exp. |
| 1 | 0.161926968 | 9 | 8.0963 | 9 | 8.0963 |
| 2 | 0.168545644 | 3 | 8.4273 | 3 | 8.4273 |
| 3 | 0.037262021 | 5 | 1.8631 | | |
| 4 | 0.162660577 | 10 | 8.1330 | 15 | 9.9961 |
| 5 | 0.015025858 | 1 | 0.7513 | | |
| 6 | 0.017927913 | 4 | 0.8964 | | |
| 7 | 0.109949741 | 3 | 5.4975 | 8 | 7.1452 |
| 8 | 0.099373226 | 5 | 4.9687 | | |
| 9 | 0.037554998 | 3 | 1.8777 | 8 | 6.8464 |
| 10 | 0.189773053 | 7 | 9.4887 | 7 | 9.4887 |
| Total | 1 | 50 | 50 | 50 | 50 |
| | $\chi^2$ | 22.5925 | | 7.0503 | |
| | df | 9 | | 6 | |
| | p-value | 0.007 | | 0.217 | |

There are problems in conducting the test in 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 in Appendix A in the supplementary material so as to illustrate the response a potential user would have when carrying out this test with this kind of data. Applying the automatic regrouping of strata with the procedure developed in this paper as introduced in the next section and the custom functions in Excel and Mathematica that we present in the supplementary material in Appendix C, the data are regrouped into 6 categories. The last two columns of Table 1 show how the categories are regrouped. Considering the 6 categories after regrouping, the sample value obtained for the $\chi^2$ statistic with 5 degrees of freedom gives a p-value equal to 0.217. It can be seen that without regrouping the categories the null hypothesis is rejected, but when the custom functions regroup to meet the minimum size requirement it is not rejected. If the minimum size requirement for validating the test is not

taken into account, the results could be wrong and, in this case, opposite to the case of regrouping.

After a comprehensive review of the software that can carry out this test, Table A1 in Appendix A in the supplementary material summarizes whether selected software packages can be used for statistical purposes to check whether Pearson's $\chi^2$ goodness-of-fit test, or at least whether 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 statistic (1) is computed. Many computer programs 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 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 it 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 procedure starts by analysing the size of the expected value in the first category. If it finds a category that does not reach the minimum size required, **minExp**, then it will be added to the next category. It follows the same regrouping criterion down to the end, where if the last category does not have the minimum size it will be added to the nearest one where the condition holds. The method then counts the number of elements in each category and calls up **chi2** to compute the chi-square test statistic value.

Therefore, there is 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. Automatic regrouping of strata: the procedure

The automatic regrouping of categories or strata is a sequential procedure that starts with an individual analysis of the size of each stratum. Before the procedure is applied, one must know the observed and expected values to be compared in the test. The expected values can either come from a fully specified population distribution or from a theoretical distribution with unknown parameters to be estimated from the observed sample values. 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, for example, geographical grouping to follow 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 centres, and online submissions of tax return forms within the deadline. In particular, the automatic strata regrouping procedure proposed analyses 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 the stratification procedure. The regrouping starts from the first category and goes down to the last one. If a category does not reach the minimum size it is added to the smallest adjacent category. If there are adjacent categories of the same size the proposed procedure will add it to the next one, the one with a larger numbering index. A flowchart of the algorithm is given in Figure 1. Three enlargements of parts of this flowchart are given in Figures 2, 3 and 4, showing the steps involved in the regrouping process on which the subsequent computation procedure is based. The main elements and the dynamics of the chart displayed in the aforementioned figures 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. 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, 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. Check whether the expected value for the first category reaches the minimum size, **min**.

5. If the expected value for the first category does not reach the minimum value and given that it does not have a previous category, its elements will be added to the second category.

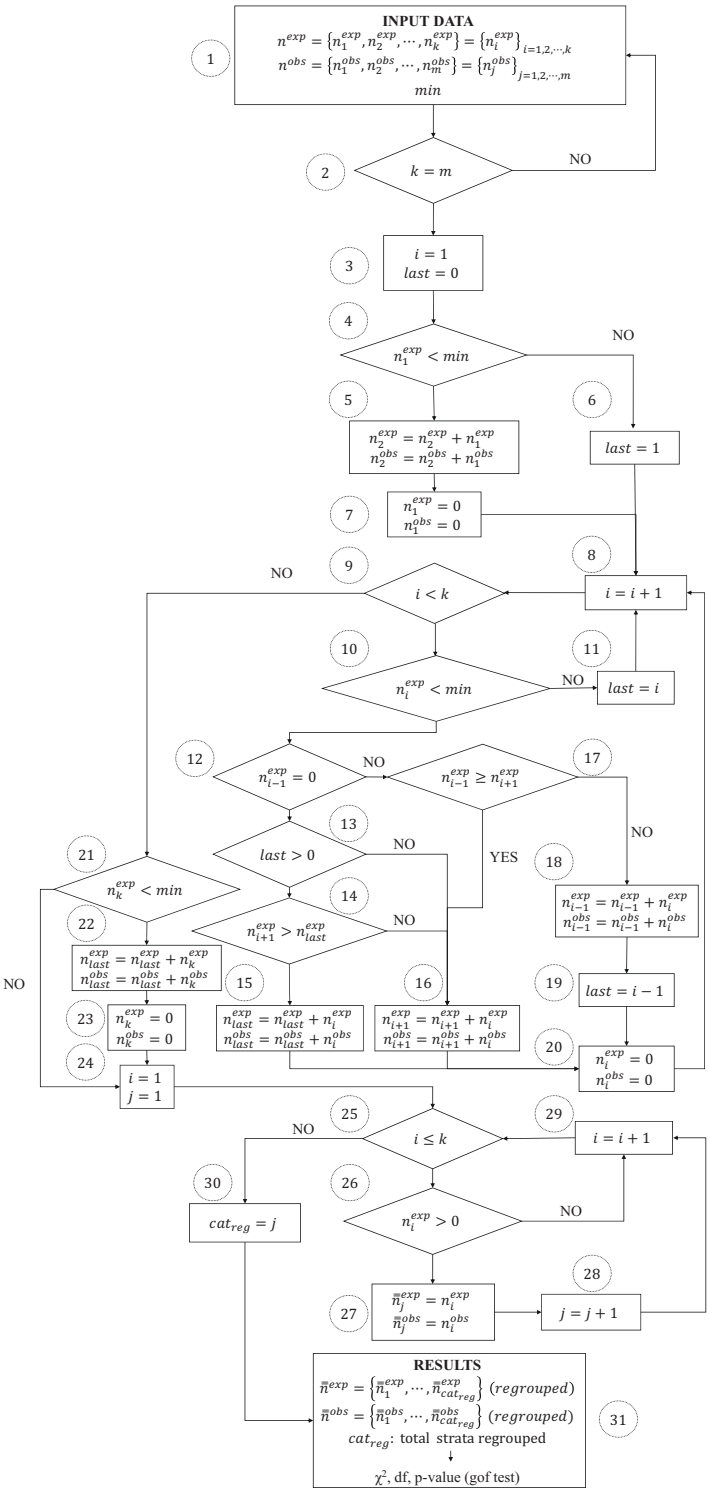


***Figure 1:*** *Flowchart. Automatic regrouping of strata.*

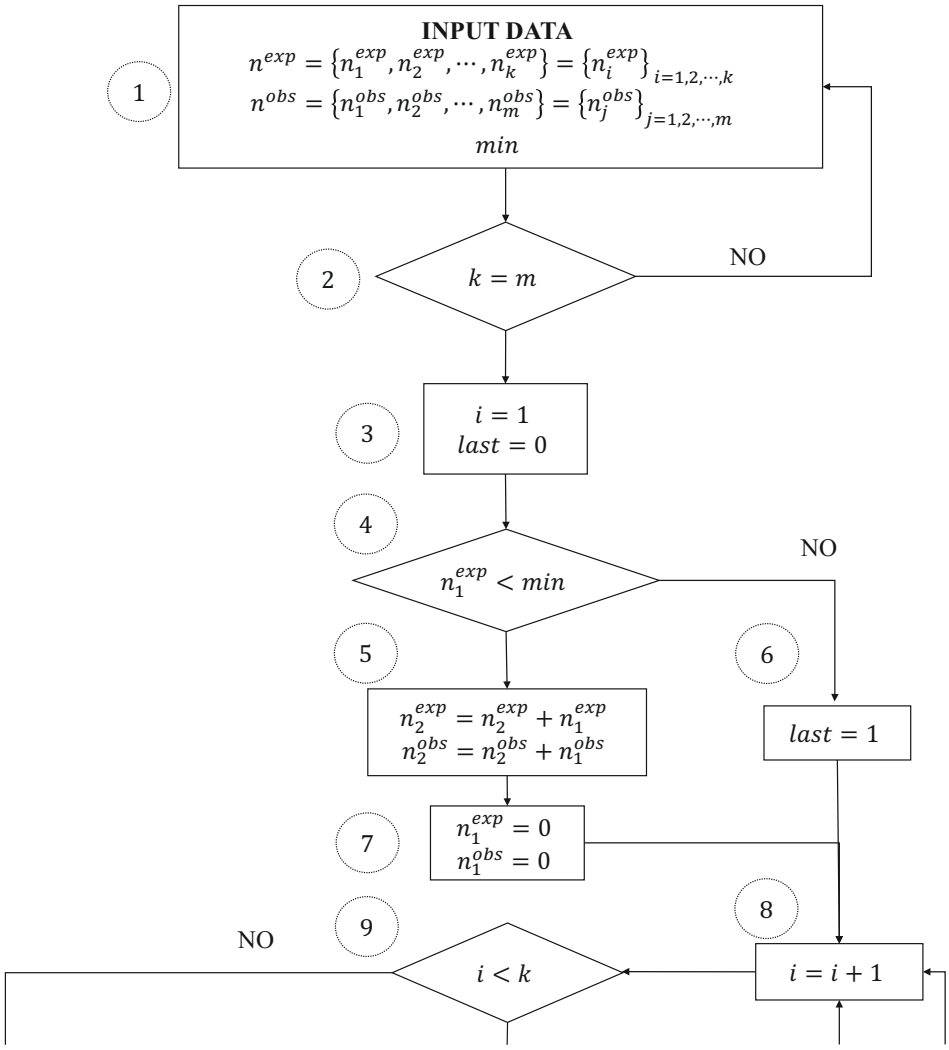


***Figure 2:*** *Flowchart: Steps 1 to 9. Automatic regrouping of strata.*

6. If the expected value for the first category reaches the minimum size, **min**, it will be stored in the variable **last**, to be, initially, the last category to reach this minimum.

7. If the expected values for the first category have been added to the second one, then the values of the first category will be initialized to zero.

8. The index *i* will increase to proceed with the analysis of the subsequent categories.

9. 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, continue with the next step. If the last stratum is reached, $i = k$, go to step 21 (see Figure 3).

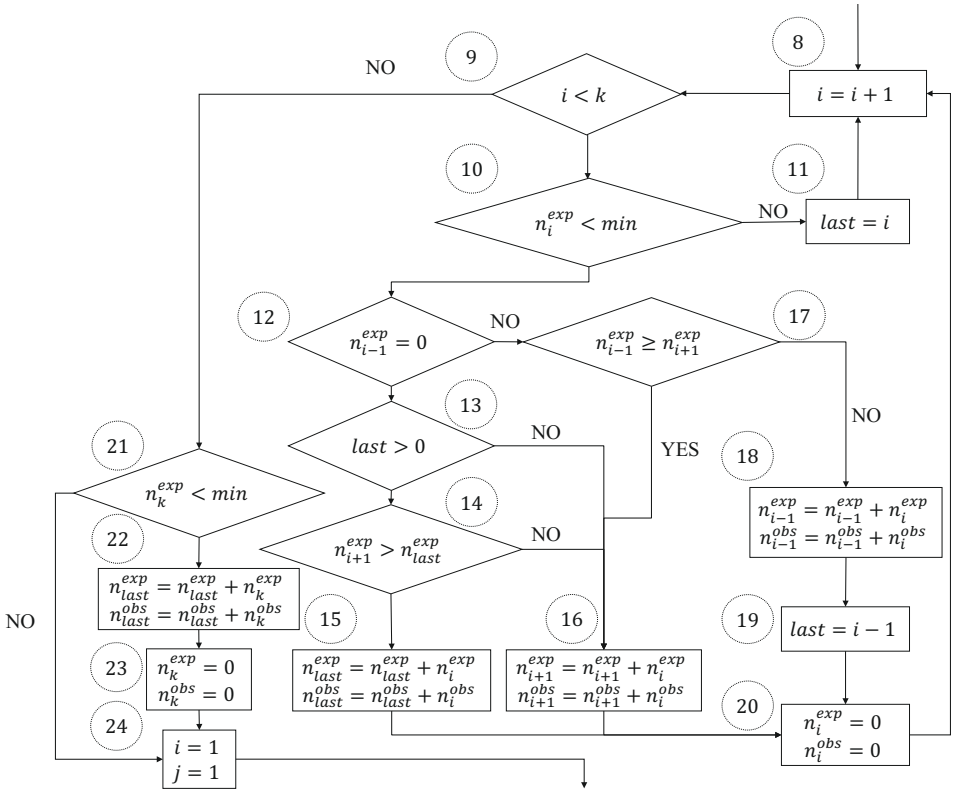


***Figure 3:*** *Flowchart: Steps 9 to 24. Automatic regrouping of strata.*

10. The expected value in stratum $i$, $n_i^{exp}$, is compared with the minimum size established at the beginning, **min**. It is worth mentioning that, except for the first category or stratum, the size of the expected value in a category to be compared with the minimum is that obtained after the loop 9-8-4-9 is performed, where step 4 is only performed for $i = 1$. In other words, it might be the result of the sum of the original value for this category and previous ones which have failed to reach the required minimum size.

11. If the expected value of a category reaches the minimum size, **min**, it is stored in the variable **last**, to be the last category to reach this minimum. Then proceed to check the next one (i.e. steps 8-9).

12. If the size of the expected value in a category does not reach the minimum, check whether the previous one is empty (i.e. it takes a value of zero).

13. If the value of a category $i$ does not reach the minimum and the immediately previous category $i-1$ is empty, check whether there is a previous non-empty category, that has a size greater than the minimum, **last**$> 0$.

14. If the value of a category $i$ does not reach the minimum, the previous one, $i-1$, is empty and there is a previous category that is not empty, **last**$> 0$, then compare the expected value of the next adjacent category, $i+1$ with the one of the previous non-empty category that reaches the minimum value; that is, the category with the index of **last**.

15. If the value of a category $i$ does not reach the minimum, the previous one, $i-1$, is empty, there is a previous category that is not empty, **last**$> 0$, and the expected value of the next adjacent category is greater than the previous non-empty one that reaches the minimum value, then the values of the category analysed, $i$, are added to the nearest previous non-empty category, **last**.

$$n_{last}^{exp} = n_{last}^{exp} + n_i^{exp}$$

$$n_{last}^{obs} = n_{last}^{obs} + n_i^{obs}$$

After that, the values of the category analysed are reset (i.e. step 20), and the next one is then analysed (i.e. steps 8-9).

16. From the second category, the values of the category analysed are added to the following one, $n_{i+1}^{exp} = n_{i+1}^{exp} + n_i^{exp}$, $n_{i+1}^{obs} = n_{i+1}^{obs} + n_i^{obs}$ when the expected value does not reach the minimum, **min**, and some of the following conditions are met:

    - The immediately previous category, already analysed, is not empty because it reached the minimum size required, but its expected value is greater than or equal to the value of the next category, $n_{i-1}^{exp} \geq n_{i+1}^{exp}$;
    - There is no previous category already analysed that meets the minimum size requirement (i.e. all are empty), so that **last** $= 0$;
    - The immediately previous category, already analysed, is empty. That is, there is a one previous category that reached an expected value equal to or greater than the minimum, but at the same time is not minor than the next category to be analysed, $n_{i+1}^{exp} \leq n_{last}^{exp}$.

17. If the expected value of the category analysed does not reach the minimum, **min**, and the previous category is not empty, compare the size of the previous category with that of the subsequent one, $n_{i-1}^{exp} \geq n_{i+1}^{exp}$.

18. If the expected value of the category analysed does not reach the minimum, **min**, and the previous category is smaller than the subsequent one, $n_{i-1}^{exp} < n_{i+1}^{exp}$ but not

empty, the values of the category analysed are added to the previous category because it is the smallest size adjacent category.

$$n_{i-1}^{exp} = n_{i-1}^{exp} + n_i^{exp}$$

$$n_{i-1}^{obs} = n_{i-1}^{obs} + n_i^{obs}$$

19. Once the values of the category analysed have been added to the previous one (in step 18) the index of that category is stored in the variable **last**$= i-1$ because it is the last one to reach the minimum value.

20. Once the values of the category analysed have been added to the previous one (in step 18), the subsequent one (in step 16) or to the one with the index **last** (in step 15), the category analysed is initialized, $n_i^{exp} = 0$, $n_i^{obs} = 0$, and the next category is then analysed.

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

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

23. After the expected and observed values of the last category, $k$, are added to the category **last** (in step 22), reset them all to zero.

24. The indexes for the categories (original and regrouped) are initialized, $i = 1$, $j = 1$.

25. Start a new loop (steps 25-29) to put together the vector of regrouped expected and observed values obtained in the previous steps. This loop is performed for all the expected values of the different strata, from the first to the last, $k$, i.e. for all $i \leq k$.

26. Check whether the accumulated expected value is greater than 0, $n_i^{exp} > 0$, which means, given what is mentioned in step 20, that it will be greater than the minimum.

27. If after regrouping the accumulated expected value of the $i$-th category is greater than 0 and, therefore, greater than the minimum, that value is assigned as the $j$-th component of a new vector of regrouped expected values, $\overline{\overline{n}}_j^{exp} = n_i^{exp}$, and the $i$-th updated observed value is assigned to the $j$-th component of the new vector of regrouped observed values $\overline{\overline{n}}_j^{obs} = n_i^{obs}$.

28. Count the number of regrouped strata put together up to this point, adding 1 to the index variable of regrouped strata in the new vectors (i.e., $j = j+1$).

29. Increment the index $i$ for the original categories of the expected values: $i = i+1$, up to the maximum, $k$.

30. Once the loop in steps 25-29 ends, the final number of regrouped categories in the new vector of expected values, $cat_{reg} = j$, is obtained.

31. The information that enables Pearson's chi-square goodness-of-fit test ($\chi^2$, df, p-value) to be carried out after the regrouping of strata is now available: $\{\overline{\overline{n}}_1^{exp}, \overline{\overline{n}}_2^{exp}, \ldots, \overline{\overline{n}}_{cat_{reg}}^{exp}\} = \{\overline{\overline{n}}_1^{obs}, \overline{\overline{n}}_2^{obs}, \ldots, \overline{\overline{n}}_{cat_{reg}}^{obs}\}$: total strata regrouped.

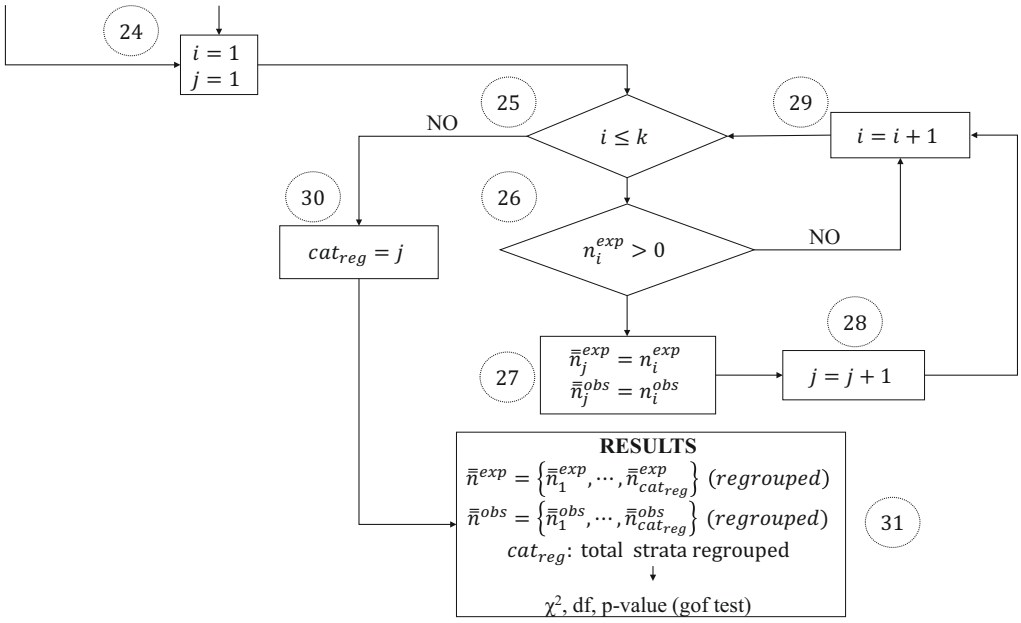


***Figure 4:*** *Flowchart: Steps 25 to 31. Automatic regrouping of strata.*

The procedure for the regrouping of strata or categories given a minimum size is written in Excel VBA. As reported by McCullough (2008), it is well known that there are quite a few shortcomings in this statistical package; however he also pointed out, as Wilkinson (1994) and Ripley (2002) claimed, that it is the most commonly used software in basic statistical calculations. This is one of the main reasons for analysing its precision (Keeling and Pavur 2011), and to provide functions that can be incorporated into the Microsoft Excel Function Library to help other users, as other authors have already done (e.g., Okeniyi and Okeniyi 2012) or, for example, to improve Excel as a useful tool for teaching (Quintela-del-Río and Francisco-Fernández 2017). In the specialized literature there is an example of using Visual Basic (Khan 2003) and its relation to Fisher's exact test (FET). This test calculates the probability value for the relationship between two dichotomous variables in a $2 \times 2$ 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 of the user friendliness of the program, its object-oriented feature and the fact that 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) pointed out, Mathematica cannot be really categorized as a statistical package, but it has complements for carrying out statistical analysis with more precision than other statistical packages. The functions are inspired by the work of Ross (2015) and Pérez-Salamero González (2015), the latter 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 $\chi^2$ test. In other words, the programming relies on the functions already available which are related to the test.

Listing 1 and Listing 2 (the latter for Mathematica) in the supplementary material in Appendix C include the code of the functions that yield the value of the $\chi^2$ statistic after automatic regrouping starting from a minimum value set by the user. The length of the code can be attributed 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 choose 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 in such a way that it selects two columns and as many rows as there were original categories, so that the user can see the regrouped categories as well as those with zero values. In the case of Mathematica, the function that returns the vectors of observed and expected values is designed to put together 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 (steps 25-29). Listing 3 in the supplementary material in Appendix C includes 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, Listing 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 we incorporate the number of categories (see Listing 5 in the supplementary material in Appendix C), the p-value for the test (Listing 6). Finally, Listing 7 shows the relevant information resulting from the regrouping procedure, such as the value of the $\chi^2$ statistic, the p-value, and the regrouped strata (observed and expected).

## 4. Simulation study

The purpose of this simulation study is to illustrate the performance of the proposed regrouping procedure on the goodness-of-fit chi-square test. The simulation study will focus on showing whether or not the proposed regrouping procedure attains the nominal significant level. We consider two different settings:

S1. The null hypothesis includes a fully specified model, so there are no parameters to be estimated.
S2. The null hypothesis includes a partially specified model in which parameters need to be estimated to compute the theoretical expected frequencies under the null hypothesis before the value of the goodness-of-fit chi-square test statistic is computed.

### *4.1. Fully Specified Population Distribution*

**Simulation 1**. The complete simulation steps are described below:

1. Six different combinations of the number of observations available, $N$, and the number of categories, $k$, are considered: A ($N = 50$, $k = 10$) , B ($N = 75$, $k = 15$), C ($N = 100$, $k = 20$), D ($N = 1000$, $k = 20$), E ($N = 500$, $k = 20$) and F ($N = 250$, $k = 20$).

2. For each combination, 5000 samples are generated from 100 different, fully specified multinomial populations under the null hypothesis, covering a wide range of possible multinomial probabilities distributions.

3. Once the 5000 samples have been generated for each combination and under each different multinomial population, we use the goodness-of-fit chi-square test statistic to test whether the data fits the theoretical distribution without regrouping. Under the null hypothesis, the chi-square statistic follows a chi-square distribution with $(k-1)$ degrees of freedom. We use three different nominal significance levels, ($\alpha$=0.10, 0.05 and 0.01). Hence, the significance levels attained are computed, corresponding to the number of times that the null hypothesis is rejected for each of the 5000 samples.

4. To assess the behaviour of the procedure proposed the same thing is done, but in this case, the categories are regrouped in those samples where the procedure proposed suggests that regrouping of some of the adjacent categories is necessary. In this case, the chi-square statistic follows a chi-square distribution with $(k-1)$ degrees of freedom under the null if it is not necessary to regroup, and with $(k^*-1)$ degrees of freedom if it is, where $k^*$ is the number of classes remaining after regrouping. Three different nominal significance levels are used, ($\alpha$=0.10, 0.05 and 0.01), and the significance levels attained are computed as before.

5. Finally, the significance levels attained for the three nominal significance levels under study are compared, without no regrouping procedure and with the regrouping procedure proposed here.

Table 2 summarizes the results of the simulation study described above. Because it is realized that the different settings mean that these results cannot really be combined, the table includes the mean and standard deviation of the significance levels attained for the 5000 simulations in each of the six $(N,k)$ combinations considered for the 100 different multinomial populations in the null hypothesis. The results shown in the table lead us to conclude that the regrouping procedure proposed provides mean attained significance levels closer to the nominal ones than those obtained by not regrouping. Moreover, standard deviations for the attained nominal significance levels are smaller when the regrouping procedure proposed is used.

***Table 2:*** *Simulation 1. Mean attained and standard deviations from nominal significance levels for the 5000 simulations in each of the six $(N,k)$ combinations considered for the 100 different multinomial populations in the null hypothesis of a fully specified population distribution for the chi-square goodness-of-fit test.*

| | | **nominal significance level** | | | | | |
|---|---|---|---|---|---|---|---|
| **populations** | | $\alpha$ **= 10%** | | $\alpha$ **= 5%** | | $\alpha$ **= 1%** | |
| | | **do not reg** | **reg** | **do not reg** | **reg** | **do not reg** | **reg** |
| **A** | mean | 0.1019 | 0.0974 | 0.0558 | 0.0484 | 0.0168 | 0.0099 |
| | st. dev. | *0.0097* | *0.0037* | *0.0103* | *0.0031* | *0.0105* | *0.0013* |
| **B** | mean | 0.1049 | 0.0980 | 0.0585 | 0.0489 | 0.0175 | 0.0101 |
| | st. dev. | *0.0092* | *0.0045* | *0.0097* | *0.0031* | *0.0077* | *0.0013* |
| **C** | mean | 0.1073 | 0.0985 | 0.0613 | 0.0491 | 0.0200 | 0.0101 |
| | st. dev. | *0.0093* | *0.0042* | *0.0088* | *0.003* | *0.0079* | *0.0014* |
| **D** | mean | 0.1021 | 0.1011 | 0.0527 | 0.0508 | 0.0119 | 0.0104 |
| | st. dev. | *0.0968* | *0.0044* | *0.0041* | *0.0035* | *0.0024* | *0.0015* |
| **E** | mean | 0.1020 | 0.0998 | 0.0532 | 0.0506 | 0.0127 | 0.0106 |
| | st. dev. | *0.1048* | *0.0036* | *0.0044* | *0.0029* | *0.0028* | *0.0016* |
| **F** | mean | 0.1027 | 0.0996 | 0.0540 | 0.0501 | 0.0137 | 0.0103 |
| | st. dev. | *0.1018* | *0.0043* | *0.0052* | *0.0031* | *0.0040* | *0.0015* |
| **ALL** | mean | **0.1036** | **0.0989** | **0.0562** | **0.0495** | **0.0157** | **0.0102** |

For the sake of brevity, Tables A2 to A7 in the supplementary material show results for only 10 selected different multinomial populations out of the 100 considered in this simulation study, where the theoretical probabilities under the null hypothesis are described at the top of the tables for the different sample sizes $N$ and numbers of categories

$k$ considered in the simulation study. As indicated above, 5000 simulations from each of these populations were simulated for different $N$ and $k$, and three different nominal significance levels were considered. Significance levels attained by using the procedure without regrouping and those obtained using the regrouping procedure proposed here are reported at the bottom of the tables for each setting. From the results reported in Tables A2 to A7 in the supplementary material, and given that the significance levels attained are very close to the nominal significance levels considered here for all the different combinations of sample sizes, population distributions, and nominal significance levels considered in the study, it can be concluded that the regrouping procedure proposed performs reasonably well compared to the results obtained without regrouping in the case of a chi-square goodness-of-fit test to a fully specified distribution.

### *4.2. Partially Specified Population Distribution*

**Simulation 2**. The complete simulation steps are described below:

1. 5000 samples are generated from a known distribution, with no loss of generality: a N(0,1) distribution, for 6 different combinations of the number of observations available, $N$, and the number of categories, $k$: A ($N = 50$, $k = 10$) , B ($N = 75$, $k = 15$), C ($N = 100$, $k = 20$), D ($N = 1000$, $k = 20$), E ($N = 500$, $k = 20$) and F ($N = 250$, $k = 20$).

2. For each sample and each setting, the mean $\mu$ and the standard deviation $\sigma$ of the normal distribution are estimated using the maximum likelihood method.

3. For the 6 different combinations of $N$ and $k$, each of the 100 different multinomial probability combinations is assigned to each of the $k$ categories.

4. The different $k$ categories (i.e. the interval limits for each class or category) in the estimated distribution $N(\hat{\mu}, \hat{\sigma}^2)$ are built up so that these intervals match the probability of belonging to this class in the estimated distribution $N(\hat{\mu}, \hat{\sigma}^2)$ with that in the corresponding multinomial population considered.

5. Once the 5000 samples have been generated for each combination and under each different multinomial populations, we use the goodness-of-fit chi-square test statistic to test wether the data fits the theoretical distribution without regrouping. Under the null hypothesis, the chi-square statistic follows a chi-square distribution with $(k-r-1) = (k-2-1) = (k-3)$ degrees of freedom. Three different nominal significance levels are used: ($\alpha$=0.10, 0.05 and 0.01). Hence, the significance levels attained are computed, corresponding to the number of times that the null hypothesis is rejected for each of the 5000 samples.

6. To assess the behaviour of the proposed procedure, the same is done, but in this case, the categories are regrouped in those samples where the procedure proposed

suggests that regrouping of some of the adjacent categories is necessary. In this case, the chi-square statistic follows a chi-square distribution with $(k-3)$ degrees of freedom under the null if it is not necessary to regroup, and with $(k^*-3)$ degrees of freedom if it is, where $k^*$ is the number of remaining classes after regrouping. Three different nominal significance levels are used: ($\alpha$=0.10, 0.05 and 0.01), and the significance levels attained are computed as before.

7. Finally, the significance levels attained are compared for the three nominal significance levels under study without and then with the regrouping procedure proposed.

Table 3 summarizes the results obtained of the simulation study described above. Because it is realized that the different settings mean that these results cannot really be combined, the table includes the mean and standard deviation of the significance levels attained for the 5000 simulations in each of the six $(N,k)$ combinations considered for the 100 different multinomial population distributions for a partially specified goodness-of-fit test of the null hypothesis of a normal distribution. The results shown in the table lead us to conclude that the regrouping procedure proposed provides mean nominal significance levels closer to the nominal ones than those obtained by not regrouping, with the exceptions of the combinations A, for $\alpha = 10\%$ and $5\%$, and B, for $\alpha = 10\%$. Moreover, standard deviations for the attained nominal significance levels are smaller when the regrouping procedure proposed is used.

For the sake of brevity, Tables A8 to A13 in the supplementary material show results for only 10 selected different multinomial probability distributions out of the 100 considered in this simulation study for the null hypothesis of a normal population distribution, where the theoretical probabilities assigned to each of the classes under the null hypothesis are described in the top part of the tables for the different sample sizes $N$ and numbers of categories $k$ considered in the simulation study. As indicated above, 5000 simulations from a standard normal population were simulated for different $N$ and $k$, and three different nominal significance levels were considered. Significance levels attained by using the procedure without and then with the regrouping procedure proposed here are reported at the bottom of the tables for each setting. From the results reported in Tables A8 to A13 in the supplementary material, and given that the significance levels attained are very close to the nominal significance levels considered here for all of the different combinations of sample sizes, population distributions, and nominal significance levels considered in the study, it can concluded that the regrouping procedure proposed performs reasonably well compared to the results obtained without regrouping in the case of a chi-square goodness-of-fit test to a partially specified distribution where parameters needed to be estimated.

***Table 3:*** *Simulation 2. Mean attained and standard deviations from nominal significance levels for the 5000 simulations in each of the six $(N,k)$ combinations considered for the 100 different multinomial populations in the null hypothesis of a partially specified normal population distribution for the chi-square goodness-of-fit test.*

| | | **nominal significance level** | | | | | |
|---|---|---|---|---|---|---|---|
| **populations** | | $\alpha$ **= 10%** | | $\alpha$ **= 5%** | | $\alpha$ **= 1%** | |
| | | **do not reg** | **reg** | **do not reg** | **reg** | **do not reg** | **reg** |
| **A** | mean | 0.1134 | 0.1378 | 0.0617 | 0.0689 | 0.0180 | 0.0135 |
| | st. dev. | *0.0097* | *0.0183* | *0.0107* | *0.0101* | *0.0114* | *0.0020* |
| **B** | mean | 0.1375 | 0.1422 | 0.0890 | 0.0862 | 0.0461 | 0.0404 |
| | st. dev. | *0.0099* | *0.0049* | *0.0110* | *0.0032* | *0.0084* | *0.0013* |
| **C** | mean | 0.1107 | 0.1087 | 0.0624 | 0.0539 | 0.0202 | 0.0106 |
| | st. dev. | *0.0087* | *0.0047* | *0.0092* | *0.0030* | *0.0087* | *0.0013* |
| **D** | mean | 0.1063 | 0.1056 | 0.0554 | 0.0536 | 0.0129 | 0.0113 |
| | st. dev. | *0.1128* | *0.0041* | *0.0053* | *0.0032* | *0.0032* | *0.0014* |
| **E** | mean | 0.1080 | 0.1063 | 0.0566 | 0.0545 | 0.0137 | 0.0116 |
| | st. dev. | *0.1080* | *0.0035* | *0.0045* | *0.0028* | *0.0031* | *0.0015* |
| **F** | mean | 0.1059 | 0.1036 | 0.0563 | 0.0526 | 0.0142 | 0.0110 |
| | st. dev. | *0.1150* | *0.0036* | *0.0054* | *0.0025* | *0.0042* | *0.0013* |
| **ALL** | mean | **0.1143** | **0.1185** | **0.0643** | **0.0624** | **0.0216** | **0.0169** |

## 5. Further illustrative examples

We use three additional examples and datasets to illustrate the use of the customized functions defined in Excel and Mathematica, where the regrouping of strata or categories could arise. In the first, the functions proposed in this paper are compared with some of the software tools described in Appendix A in the supplementary material. Some of them do not automatically regroup and others, e.g. MATLAB, do so but only at the extreme ends of the tails. The second example illustrates the use of the regrouping functions when it is necessary to estimate parameters in the theoretical distribution. The third shows the iterative use of the regrouping functions with application to analyse the Continuous Sample of Working Lives (CSWL) survey from Spain.

### 5.1. Case 1. Pearson's Illustration V

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

in the bins at the extreme ends. The null hypothesis is the fully specified population distribution with probabilities described in Pearson (1900).

***Table 4:*** *Case 1. Illustration V example. Observed and expected values are reported, as well as results for the goodness-of-fit chi-square test for fully specified distributions with no regrouping of categories and with the regrouping procedure proposed here.*

| Category | Original | | Regrouped | |
|---|---|---|---|---|
| | Observed | Expected | Observed | Expected |
| 1 | 0 | 0.18 | | |
| 2 | 3 | 0.68 | | |
| 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 | | |
| 14 | 1 | 1.06 | | |
| 15 | 0 | 0.34 | | |
| 16 | 0 | 0.10 | | |
| 17 | 0 | 0.00 | | |
| Total | 500 | 500.36 | 500 | 500.36 |
| $\chi^2$ | 11.75 | | 10.51 | |
| df | 16 | | 9 | |
| p-value | 0.101 | | 0.31083538 | |
| Source: Own work based on Pearson (1900) | | | | |

Pearson (1900) considered there to be 17 categories, though the expected value of the last one is zero. Taking into account all the strata and with no regrouping, the value of the $\chi^2$ test statistic compared to a chi-square distribution with 16 degrees of freedom results in a p-value of 0.101. Moreover, the functions defined in Excel and Mathematica, presented in the supplementary material included in Appendix 2, regroup them into 10 categories. The last two columns of Table 4 show how the proposed functions regroup

***Table 5:*** *Case 2. Observed and expected values are reported, as well as results for the goodness-of-fit chi-square test for partially specified distributions with no regrouping of categories and with the regrouping procedure proposed here. The null hypothesis states that the population follows a partially specified normal distribution.*

| Categ. | pexp | expected | ac prob exp | levels | Obs | Obs. Reg. | Exp. reg. |
|---|---|---|---|---|---|---|---|
| 1 | 0.161926968 | 8.09634839 | 0.161926968 | −0.7747845 | 9 | 9 | 8.09634839 |
| 2 | 0.168545644 | 8.42728221 | 0.330472612 | −0.31307327 | 8 | 8 | 8.42728221 |
| 3 | 0.037262021 | 1.86310107 | 0.367734633 | −0.22818199 | 0 | | |
| 4 | 0.162660577 | 8.13302885 | 0.530395211 | 0.12075745 | 10 | 10 | 9.99612992 |
| 5 | 0.015025858 | 0.75129289 | 0.545421068 | 0.152639283 | 1 | | |
| 6 | 0.017927913 | 0.89639567 | 0.563348982 | 0.190863633 | 4 | | |
| 7 | 0.109949741 | 5.49748705 | 0.673298723 | 0.434859108 | 3 | 8 | 7.14517561 |
| 8 | 0.099373226 | 4.96866129 | 0.772671949 | 0.68648864 | 5 | | |
| 9 | 0.037554998 | 1.8777499 | 0.810226947 | 0.796917803 | 3 | 8 | 6.84641119 |
| 10 | 0.189773053 | 9.48865267 | 1 | 0.796917803 | 7 | 7 | 9.48865267 |
| | | | | p-value do not reg. | 0.028 | p-value reg. | 0.7839 |

the original categories. Considering the 10 categories resulting after regrouping, the value of the $\chi^2$ test statistic, when compared to a chi-square distribution with 9 degrees of freedom results in a p-value of 0.311. The problems a potential user would have when using the different software tools available for the analysis of this dataset are outlined in Appendix A in the supplementary material.

### 5.2. Case 2. Example of a partially specified population distribution

This example illustrates the use of the regrouping functions once the parameters of a partially specified theoretical distribution have been estimated by the maximum likelihood method using the sample data provided in Table A14 in the supplementary material. The example is based on the second simulation study described in Section 4. The null hypothesis states that the population follows a partially specified normal distribution and its parameters, its mean and standard deviation, are unknown and must therefore be estimated from the values reported in Table A14 in the supplementary material. Parameters are estimated by maximum likelihood, using the **fitdistrib** function in the MASS library in R. Therefore, the null hypothesis states that the data follows a $N(\widehat{\mu},\widehat{\sigma}) = N(0.056497994, 0.842599379)$ distribution. To test this hypothesis and obtain the goodness-of-fit chi-square test statistic for partially specified distributions, there are originally $k = 10$ categories and $(k-3)$ degrees of freedom for test statistic chi-square distribution in the case of not regrouping. In the case of regrouping, the degrees of freedom are $(k^*-3)$, where $k^*$ is the number of remaining categories after the regrouping procedure proposed is applied. In order to force the necessity for regrouping, different probabilities are randomly assigned to the $k = 10$ categories used for this test. Given the estimated parameters and the probabilities assigned to each category, we obtain the interval limits for these categories by using the procedure previously described in the Simulation 2 settings. Table 5 reports the information required to perform the test and the resulting p-values obtained with and without regrouping. Under the regrouping procedure proposed the null hypothesis is not rejected, but it is clearly rejected if there is no regrouping, at least at the 10% and 5% significance levels.

### 5.3. Case 3. Example with the Continuous Sample of Working Lives dataset

This example illustrates the iterative use of the proposed regrouping functions. The $\chi^2$ test statistic value 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 González 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 González, Regúlez-Castillo and Vidal-Meliá (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 some 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 González et al. (2017) obtain from the CSWL for the 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 as of December 31st (INSS 2014), which means that the relative expected frequencies are also known, and hence so are the expected values (i.e. this is a fully specified population distribution test setting). Table 6 reports the observed values from the CSWL and the expected values from the theoretical population under the null hypothesis, along with the corresponding fully specified chi-square goodness-of-fit test results with and without regrouping. From the results reported in Table 6, we conclude that the null hypothesis is clearly rejected whether regrouping is performed or not. That is, the null hypothesis is clearly rejected in the case of automatic regrouping and also in the case of no regrouping of strata. The **chisq.test** function written in Excel is used for the regrouping procedure proposed. Moreover, the fit of the sample to the population could be improved, since the null hypothesis is rejected, and 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, which is one of the main objectives for practitioners in the area.

To further show the utility of the customized functions used iteratively which enable the $\chi^2$ 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 non rejection of the null hypothesis of the assumed theoretical distribution for 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 B in the supplementary material. It is implemented in Excel by using the functions defined in the supplementary material in Appendix C, which allows for an automatic regrouping procedure. The problem is solved by using the **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 (see Appendix B).

***Table 6:*** *CSWL 2013. Permanent Disability: Males. Observed and expected values are reported, as well as results for the goodness-of-fit chi-square test for completely specified distributions without regrouping categories and with the regrouping procedure proposed here.*

| Age Category | Original | | Regrouped | |
|---|---|---|---|---|
| | Observed | Expected | Observed | Expected |
| 15-19 | 0 | 0.04 | | |
| 20-24 | 29 | 30.04 | 29 | 30.08 |
| 25-29 | 198 | 195.33 | 198 | 195.33 |
| 30-34 | 606 | 581.48 | 606 | 581.48 |
| 35-39 | 1,201 | 1,203.73 | 1,201 | 1,203.73 |
| 40-44 | 2,014 | 1,982.02 | 2,014 | 1,982.02 |
| 45-49 | 3,106 | 3,050.46 | 3,106 | 3,050.46 |
| 50-54 | 4,281 | 4,230.30 | 4,281 | 4,230.30 |
| 55-59 | 5,710 | 5,706.36 | 5,710 | 5,706.36 |
| 60-64 | 7,151 | 7,269.83 | 7,151 | 7,269.83 |
| 65-69 | 3 | 58.48 | 3 | 58.48 |
| 70-74 | 6 | 3.28 | | |
| 75-79 | 7 | 4.28 | 13 | 7.56 |
| 80-84 | 14 | 10.88 | 14 | 10.88 |
| $\geq 85$ | 17 | 16.48 | 17 | 16.48 |
| Total | 24,343 | 24,343 | 24,343 | 24,343 |
| $\chi^2$ | 62.76 | | 62.66 | |
| df | 14 | | 12 | |
| p-value | p-value$<$0.0001 | | p-value$<$0.0001 | |
| Source: Own work based on Pérez-Salamero González et al. (2017) | | | | |

Accuracy in compliance with 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 $100,000$ subproblems are solved, a non-integer solution is reached ("*Solver found a probability of reaching a global solution*"). The solution is then rounded and it is finally checked that the one obtained is contained in the original sample. Constraint **[2.]** in Appendix B, related to the improvement of the goodness-of-fit, is not satisfied by a small error of 0.000000722, the difference between the sample value of the test and the critical value at the 5% significance level, with a reported p-value of 0.0499993. The emergence of this solution, with no attention paid to the minimum size requirement for the strata, is due to the functions defined in the sup-

plementary material in Appendix C. 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 regroup exogenously in the procedure within each iteration.

***Table 7:*** *Subsample from the CSWL 2013. Permanent disability: Males. Observed and expected values are reported, as well as results for the goodness-of-fit chi-square test for completely specified distributions without regrouping categories and with the regrouping procedure proposed here. In addition, results for the subsample obtained with the proposed algorithm are also reported.*

| Age Category | Original sample CSWL | | Subsample (before rounding) | | | |
|---|---|---|---|---|---|---|
| | Observed | Observed (regrouped) | Observed | Expected | Observed (regrouped) | Expected (regrouped) |
| 15-19 | 0 | | 0 | 0.02 | | |
| 20-24 | 29 | 29 | 13.03 | 12.98 | 13.03 | 13.00 |
| 25-29 | 198 | 198 | 84.59 | 84.41 | 84.59 | 84.41 |
| 30-34 | 606 | 606 | 251.80 | 251.27 | 251.80 | 251.27 |
| 35-39 | 1,201 | 1,201 | 521.25 | 520.15 | 521.25 | 520.15 |
| 40-44 | 2,014 | 2,014 | 858.27 | 856.46 | 858.27 | 856.46 |
| 45-49 | 3,106 | 3,106 | 1,320.94 | 1,318.14 | 1,320.94 | 1,318.14 |
| 50-54 | 4,281 | 4,281 | 1,831.85 | 1,827.97 | 1,831.85 | 1,827.97 |
| 55-59 | 5,710 | 5,710 | 2,471.03 | 2,465.80 | 2,471.03 | 2,465.80 |
| 60-64 | 7,151 | 7,151 | 3,148.06 | 3,141.39 | 3,148.06 | 3,141.39 |
| 65-69 | 3 | 3 | 3 | 25.27 | 3 | 25.27 |
| 70-74 | 6 | | 2.25 | 1.42 | | |
| 75-79 | 7 | 13 | 0.25 | 1.85 | | |
| 80-84 | 14 | 14 | 5.48 | 4.70 | 7.99 | 7.97 |
| $\geq 85$ | 17 | 17 | 7.14 | 7.12 | 7.14 | 7.12 |
| Total | 24,343 | 24,343 | 10,518.94 | 10,518.94 | 10,518.94 | 10,518.94 |
| $\chi^2$ | 62.76 | 62.66 | 21.69 | | 19.68 | |
| df | 14 | 12 | 14 | | 11 | |
| p-value | $p < 0.0001$ | $p < 0.0001$ | 0.0851783 | | 0.0499993 | |
| Source: Own work based on Pérez-Salamero González et al. (2017) | | | | | | |

The results of the optimization process and the size of the strata associated with the solution obtained are presented in Table 7. The first two columns in Table 7 correspond to the first and third columns of Table 6, and we report them back in order to improve the comparison between the original sample and the subsample obtained. The last four columns in Table 7 have the same structure as the ones shown in Tables 1, 4, 5 and 6.

Table 7 shows that the p-value of 0.085 obtained for the $\chi^2$ goodness-of-fit test in the subsample with no regrouping of strata results in no rejection of the null hypothesis, whereas the p-value of 0.04999928 obtained after regrouping is at the limit of rejection of the null hypothesis, both at the 5% significance level.

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

## 6. Summary, conclusions and further research

In empirical studies where Pearson's goodness-of-fit $\chi^2$ test is conducted, it is a common practice to regroup strata to attain a minimum size of expected frequencies for the test to be valid and its conclusions reliable. In general, after a comprehensive review of 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 such a feature were available. Having such automatic regrouping available in other packages would be of great help to researchers in many areas, such as social sciences, biomedical and health sciences, and others where this test is usually required 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, as in SSJ (a Java library for stochastic simulation).

A simulation study shows that the regrouping functions proposed in this paper work reasonably well compared to the test without regrouping. We find that the nominal significance levels attained with regrouping are suitable and slightly better than those obtained without regrouping. They guarantee that the hypotheses of the minimum size are satisfied, reducing the risk of a wrong conclusion on the goodness-of-fit chi-square test. The customized functions developed here have the advantage of being easier to implement than SSJ in an iterative process, where the test statistic must be calculated and the regrouping carried out in each iteration. Moreover, they offer an alternative way of regrouping that solves the asymmetry problem in the test results. This type 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 enable Pearson's goodness-of-fit chi-square test to be carried out with the possibility of regrouping categories, which we believe is quite a major improvement on the current software available for basic statistical analysis, both in the case of the most widely used program, Excel, and other more precise packages

such as Mathematica. We also believe that these proposals 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 and Larose, 2014), for example, at the instance selection and association analysis stages, among others.

Finally, based on the proposals included and results reported in 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, R, etc, in order to be able to integrate them into the numerical resolution of problems of this type. It would also be interesting to make the regrouping process automatic, but based on other, more general criteria, such as, for example, sample size or number of categories, and to explore alternative ways of regrouping. This would require analysing the effect of the different regrouping proposals on the goodness-of-fit chi-square test results for different sample sizes, number of categories and theoretical distributions under study. This is out of the current scope of this paper, but could be the objective of future research.

## Acknowledgements

The authors gratefully acknowledge financial support from Ministerio de Economía y Competitividad (Spain), Agencia Estatal de Investigación (AEI), and the European Regional Development Fund (ERDF), under research grants ECO2015-65826-P (AEI/ERDF, EU) and MTM2016-74931-P (AEI/ERDF, EU) and from the Department of Education of the Basque Government (UPV/EHU MacLab Research Group and UPV/EHU Econometrics Research Group) under research grants IT 793-13 and IT-642-13, respectively. The authors wish to thank the editor and two anonymous referees for providing thoughtful comments and suggestions which have led to substantial improvement in the presentation of the material in this paper. They also would like to thank Jose M. Pavía, Miguel Angel García Pérez and Fernando Tusell for their comments and suggestions, and Christopher G. Pellow for his help with the English. Any errors are entirely due to the authors.

## References

Agresti, A. (2002). *Categorical Data Analysis* (2nd edition). Wiley, New York.

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

Bartholomew, D.J., Knott, M. and Moustaki, I. (2011). *Latent Variable Models and Factor Analysis* (3rd edition). Wiley, New York.

Bishop, Y.M.M., Fienberg, S.E. and 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, 167–172.

Cai, L., Maydeu-Olivares, A., Coffman, D.L. and Thissen, D. (2006). Limited-information goodness-of-fit testing of item response theory models for sparse 2p tables. *British Journal of Mathematical and Statistical Psychology*, 59, 173–194.

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

Cochran, W.G. (1952). The $\chi^2$ test of goodness-of-fit. *The Annals of Mathematical Statistics*, 23, 315–345.

Collins, L.M., Fidler, P.L., Wugalter, S.E. and Long, J. (1993). Goodness-of-fit testing for latent class models. *Multivariate Behavioral Research*, 28, 375–389.

Delucchi, K.L. (1983). The use and misuse of chi-square: Lewis and Burke revisited. *Psychological Bulletin*, 94, 166–176.

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

Fienberg, S.E. (2006). Log-linear models in contingency tables. In *Encyclopedia of Statistical Sciences*. Wiley, New York.

Fisher, R.A. (1935). The logic of inductive inference. *Journal of the Royal Statistical Society*, 98, 39–54.

García Pérez, M.A. and Nuñez-Antón, V. (2009). Accuracy of power-divergence statistics for testing independence and homogeneity in two-way contingency tables. *Communications in Statistics - Simulation and Computation*, 38, 503–512.

Goodman, L.A. (1974). Exploratory latent structures analysis using both identifiable and unidentifiable models. *Biometrika*, 61, 215–231.

Grafstörm, A. and Schelin, L. (2014). How to select representative samples. *Scandinavian Journal of Statistics*, 41, 277–290.

Haviland, M.G. (1990). Yates´ s correction for continuity and the analysis of $2 \times 2$ contingency-tables. *Statistics in Medicine*, 9, 363–367.

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

Hosmer, D.W., Hosmer, T., Le Cessie, S. and Lemeshow, S. (1997). A comparison of goodness-of-fit tests for the logistic regression model. *Statistics in Medicine*, 16, 965–980.

Hosmer, D.W. and Lemeshow, S. (2000). *Applied Logistic Regression*. Wiley, New York.

INSS (2014). Informe Estadístico 2013. Secretaría de Estado de Seguridad Social. Ministerio de Empleo y Seguridad Social, MESS. Madrid, Spain.

Keeling, K.B. and Pavur, R.J. (2011). Statistical accuracy of spreadsheet software. *The American Statistician*, 65, 265–273.

Khan, H.A. (2003). A visual basic software for computing Fisher´s exact probability. *Journal of Statistical Software*, 8, 1–7.

Kroonenberg, P.M. and Verbeek, A. (2018). 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*, 72, 175–183.

Kruskal, W. and Mosteller, F. (1979a). Representative sampling, I. *International Statistical Review*, 47, 13–24.

Kruskal, W. and Mosteller, F. (1979b). Representative sampling, II: scientific literature, excludind statistics. *International Statistical Review*, 47, 111–127.

Kruskal, W. and Mosteller, F. (1979c). Representative sampling, III: the current statistical literature. *International Statistical Review*, 47, 245–265.

Kruskal, W. and Mosteller, F. (1980). Representative sampling, IV: The History of the Concept in Statistics, 1895-1939. *International Statistical Review*, 48, 169–195.

Larose, D.T. and Larose, C.D. (2014). *Discovering Knowledge in Data: An Introduction to Data Mining*. Wiley, New York.

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

Lewis, D. and Burke, C.J. (1949). The use and misuse of chi-square. *Psychological Bulletin*, 46, 433–489.

Lin, J.J., Chang, C.H. and 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, 438–458.

Lydersen, S., Fagerland, M.W. and Laake, P. (2009). Tutorial in biostatistics. Recommended tests for association in 2x2 tables. *Statistics in Medicine*, 28, 1159–1175.

Marsaglia, G. (2003). Random number generators. *Journal of Modern Applied Statistical Methods*, 2, 2–13.

McCullough, B.D. (2000). The accuracy of Mathematica 4 as a statistical package. *Computational Statistics*, 15, 279–299.

McCullough, B.D. (2008). Special section on Microsoft Excel 2007. *Computational Statistics and Data Analysis*, 52, 4568–4569.

Mehta, C.R. and Patel, N.R. (1983). A network algorithm for performing Fisher's exact test in r×c contingency tables. *Journal of the American Statistical Association*, 78, 427–434.

MESS (2017). La Muestra Continua de Vidas Laborales. Guía del contenido. Estadísticas, Presupuestos y Estudios. Estadísticas. Secretaría de Estado de Seguridad Social. Ministerio de Empleo y Seguridad Social, MESS. Madrid, Spain.

Moore, D.S. (1986). Tests of chi-squared type. In *Goodness-of-fit Techniques* (R. D'Agostino and M. Stephens, eds.). Marcel Dekker, New York, 63–95.

Okeniyi, J.O. and 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, 1727–1741.

Omair, A. (2014). Sample size estimation and sampling techniques for selecting a representative sample. *Journal of Health Specialties*, 2, 142–147.

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, 157–175.

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. Unpublished Ph.D. Thesis. Universitat de Valéncia, Spain.

Pérez-Salamero González, J.M., Regúlez-Castillo, M. and 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, 67–130

Pérez-Salamero González, J.M., Regúlez-Castillo, M. and Vidal-Meliá, C. (2017). The continuous sample of working lives: improving its representativeness. *SERIEs. Journal of the Spanish Economic Association*, 8, 43–95.

Quintela-del-Río, A. and Francisco-Fernández, M. (2017). Excel templates: a helpful tool for teaching statistics. *The American Statistician*, 71, 317–325.

Ramsey, C.A. and Hewitt, A.D. (2005). A methodology for assessing sample representativeness. *Environmental Forensics*, 6, 71–75.

Ripley, B.D. (2002). Statistical methods need software: a view of statistical computing. *Opening lecture - Royal Statistical Society*, Plymouth.

Ross, A. (2015). Probability or statistics-permorming a chi-square goodness-of-fit test. *Mathematical Stack Exchange*.

Tollenaar, N. and 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, 271–288.

Tsang, W.W. and Cheng, K.H. (2006). The chi-square test when the expected frequencies are less than 5. In *COMPSTAT 2006 - Proceedings in Computational Statistics* (A. Rizzi and M. Vichi, eds.). Physica Verlag - Springer, Heidelberg, 1583–1589.

Wickens, T.D. (1989). *Multiway Contingency Tables Analysis for the Social Sciences*. Hillsdale, NJ: Erlbaum.

Wilkinson, L. (1994). Practical guidelines for testing statistical software. In *Computational Statistics: Papers Collected on the Occasion of the 25th Conference on Statistical Computing at Schloss Reisensburg* (P. Dirschedl and R. Ostermann, eds.). Physica Verlag - Springer, Heidelberg, 1–16.

Yates, F. (1934). Contingency tables involving small numbers and the $\chi^2$ test. *Supplement to the Journal of the Royal Statistical Society*, 1, 217–235.