RECyT
Year 27 / Nº 44 / 2025 /
DOI: https://doi.org/10.36995/j.recyt.2025.44.001

Strengthening statistical tests in low-frequency and fuzzy number contingency tables through linear scaling

Fortaleciendo pruebas estadísticas en tabla de contingencia con baja frecuencia y números borrosos a través del escalado lineal

Matilde I., Césari[1, *]; Santiago, Pérez[1]

1- Grupo Regional de Investigación y Desarrollo en Ecosistemas de Conocimiento (GiDeCo). Universidad Tecnológica Nacional. Mendoza, Argentina.
* E-mail: matilde.cesari@frm.utn.edu.ar

Received: 12/11/2024; Accepted: 18/09/2025

Abstract
This study introduces an innovative methodology aimed at enhancing the application of statistical tests in contingency tables with low expected cell frequencies. By means of a linear scaling technique, we address the limitations of traditional tests—such as the Chi-Square and Fisher's Exact Test—when dealing with small values and fuzzy data, thereby facilitating a robust statistical analysis adapted to conditions of uncertainty. Classical contingency table analysis presents constraints when cell values are extremely small, particularly when decimal values fall below one. In such contexts, methods that rely on these tables, including the Chi-Square Test and Fisher's Exact Test (FET), may prove inadequate for appropriately managing the associated uncertainty, especially when the degree of imprecision within the data is non-negligible. This study develops: (1) an innovative adaptation of contingency tables that enhances the robustness of statistical tests under suboptimal sampling conditions, and (2) a computational tool that automates its implementation. Moreover, examples of scalability are discussed as a resource applicable to any contingency table facing the challenge of small values, as well as the effective use of information embedded in fuzzy or imprecise data across diverse research domains.

Keywords: Contingency table, Linear scaling, Statistical tests.

Resumen
Este trabajo presenta una metodología innovadora para fortalecer la aplicación de pruebas estadísticas en tablas de contingencia con baja frecuencia esperada en las celdas. Con esta técnica de escalado lineal, abordamos las limitaciones de pruebas tradicionales, como Chi-Cuadrado y Fisher, al enfrentar valores pequeños y datos borrosos, facilitando un análisis estadístico robusto y adaptado a condiciones de incertidumbre. El análisis clásico de Tabla de Contingencias presenta una limitante para valores muy pequeños, en las celdas de las tablas, con valores decimales menores a uno. En este contexto, métodos que involucran su utilización, como la Prueba de Chi-Cuadrado y Prueba Exacta de Fisher (FET) pueden no ser apropiados para manejar adecuadamente la incertidumbre asociada, y en especial, cuando el grado de la misma que poseen los datos no es despreciable.
En este trabajo se desarrolló: 1) una adaptación innovadora de las Tablas de contingencia, que mejora la robustez de las pruebas estadísticas bajo condiciones de muestra subóptima y 2) una herramienta informática que automatiza su aplicación. Asimismo, se discuten ejemplos de la escalabilidad como herramienta para cualquier Tabla de contingencia que enfrente el desafío de valores pequeños, así como también el aprovechamiento eficaz de la información contenida en datos borrosos o imprecisos en una variedad de campos de investigación.

Palabras clave: Escalado lineal, Pruebas estadísticas, Tabla de contingencia.

1 INTRODUCTION

In numerous engineering disciplines, contingency tables employed in research are not exclusively based on integer frequencies; rather, they often incorporate very small decimal values, and even values expressed in scientific notation, due to the precision of measurement instruments and the nature of the phenomena under study. These measurements, although minimal, are highly valuable given the richness of the information they provide and the considerable time and cost associated with their collection. Nevertheless, conventional statistical analysis of such tables presents challenges, as methods like the Chi-Square Test [1–4] require higher frequencies to validate their assumptions of independence. This issue is exacerbated in situations where the small magnitude and precision of the data are critical for

accurate interpretation of results. In this context, the linear scaling proposed herein offers an innovative solution, enabling these small decimal values to be transformed into an expanded and more manageable scale without compromising the integrity of the original data. This transformation facilitates the application of robust statistical tests capable of effectively handling measurement precision, thereby allowing engineers and scientists to fully leverage the potential of their data to understand and explain complex phenomena.

The Fisher's Exact Test (FET) [5–8] is a widely used non-parametric alternative to the Chi-Square Test for small datasets, particularly useful when one or more cells in a contingency table have expected frequencies below five. Its primary objective is to test the null hypothesis that the observed dichotomous frequency distributions are associated, against the alternative hypothesis that they are independent. FET is especially effective in analyzing small datasets, notably when at least one cell in the contingency table has an expected frequency under five.

Small datasets and sparse cells may compromise the validity of the Chi-Square Test, which relies on asymptotic normality. Since FET is a non-parametric test, this assumption does not apply. FET calculates the exact probability (p-value) of observing the given table or a more extreme distribution of the data. This probability is determined by evaluating all possible rearrangements of the table (in the direction of the alternative hypothesis).

Despite their utility, both the FET and Chi-Square tests [9] exhibit limitations in adequately addressing the uncertainty inherent in low-frequency data, which may lead to erroneous statistical conclusions or diminished test power. Specifically, when comparing proportions of a categorical outcome across different independent groups, various statistical tests are considered, including the Chi-Square Test, FET, and the Mann–Whitney Test [10–12].

The Chi-Square Test and FET are suitable for evaluating the independence between two variables when the comparison groups are independent and uncorrelated. The Chi-Square Test applies an approximation assuming a large sample size, whereas the Fisher's Exact Test performs an exact procedure, particularly for small samples.

This issue is magnified in research settings, where data often contain a degree of uncertainty or vagueness that cannot be disregarded. One example of these challenges is fuzzy data analysis [13–15], where values are not classified into discrete categories but are instead assigned degrees of membership, reflecting their uncertainty.

Transforming original data into fuzzy sets introduces a deliberate level of imprecision to capture the uncertainty inherent in certain measurements or subjective perceptions. This approach, which employs fuzzy logic and fuzzy variables, enables a more nuanced and realistic analysis of data, especially in contexts where absolute precision is unattainable or does not accurately reflect the nature of the observed phenomena. However, when working with possibility values derived from this fuzzy transformation, traditional statistical analysis faces additional complications.

Contingency tables based on these possibility values tend to contain inherently small numbers, posing a challenge to conventional association tests. The solution proposed in this article is to linearly scale these possibility values from the 0–1 range to a broader 0–100 range. This adjustment, although conceptually simple, is novel in its capacity to adapt contingency tables derived from fuzzy data for analysis using traditional statistical tests.

This study proposes an innovative adaptation of contingency tables that enhances the robustness of statistical tests under suboptimal sampling conditions. The central technique of this approach is the linear scaling of possibility values from the 0–1 range to an expanded 0–100 range.

This adjustment not only facilitates compliance with the statistical assumptions required for tests such as Chi-Square and Fisher's Exact Test, but also optimizes the interpretation and detection of significant patterns that might otherwise remain undetected due to the high variability and small magnitude of the original values.

In light of the above, and with the aim of contributing to a clearer understanding of the definition and scope of this pioneering methodology, Section 2 presents the Methodology, detailing contextual examples; Section 3 explores a more complex case study applying the methodology; Section 4 addresses the use of the methodology in tables with fuzzy numbers; Section 5 provides a detailed analysis of results and the advantages of using simple correspondence in scaled contingency tables; and finally, Section 6 summarizes the conclusions derived from this proposal and experimental study.

2 METHODOLOGY

2.1 Overview

The software developed in this study enables automated linear scaling, adapting to various data formats and providing an accessible interface for researchers across different fields. Its capacity to accurately scale low values in contingency tables enhances the compatibility of these data with tests

such as Chi-Square and Fisher's Exact Test, without compromising the integrity of the original proportions.

The software developed in this study, available at GitHub [21], enables automated linear scaling, adapting to various data formats and providing an accessible interface for researchers across different fields. Its capacity to accurately scale low values in contingency tables enhances the compatibility of these data with tests such as Chi-Square and Fisher's Exact Test, without compromising the integrity of the original proportions. A live version of the application can be accessed on the server at DaFu App [22].

In the present study, a computer tool has been developed that automates the linear scaling process, transforming this approach into a practical and efficient solution for researchers.

The software is designed to accept data inputs in various formats, automatically analyze the distribution of values, and apply linear scaling from 0 to 100 to each value in the contingency table.

This automation includes determining the minimum and maximum values within the dataset, as well as calculating the slope (M) and the intercept (B) required for the linear transformation, according to the formulas presented in the study.

Once scaled, the data are restructured into new contingency tables that preserve the original proportionality while amplifying the values, thereby facilitating the application of statistical tests that require specific assumptions regarding frequency size.

This tool accelerates data preparation and ensures precision and consistency in the application of linear scaling, which is crucial for the validity of subsequent statistical analyses.

Detailed analyses and illustrative examples are proposed to demonstrate the usefulness of the linear scaling of values as a methodological resource to improve the interpretation of contingency tables and provide a solid basis for decision-making. The proposal consists of scaling the values of the original contingency table, which range from 0 to 1, obtaining an equivalent table through the linear normalization method for values from 0 to 100 and subsequently deriving results from it by applying statistical tests such as Chi-Square or Fisher's Exact Test. Formally, this process involves selecting the optimal maximum and minimum; in this case, the maximum is represented by a value of one (a) and the minimum by a value of zero (b). The remaining values are then adjusted using the linear equation:

$$y = M \cdot x + B \quad (Equation\ 1)$$

Where M = slope and B = intercept.

The parameters M and B are calculated according to the maximum and minimum values of the attribute to be normalized (Equation 2). Finally, the standardized values are obtained from MMM and BBB using Equation (3).

$$M = \frac{a - b}{maximum - minimum} * B = (y - M \cdot x) = (a - M * maximum) \quad (Equation 2)$$

The parameters a and b correspond to the maximum and minimum, respectively, of the new value scale to which the data are to be standardized.

$$NormalizedValue = M * ObservedValue + B \quad (Equation\ 3)$$

2.2 Simple Example

Let us consider a simple example involving a binary contingency table with 0/1 values. In Table 1, on the left, we observe the original data, and on the right, the table linearly scaled to values ranging from 0 to 100. A simple correspondence analysis enables the examination of both tables and confirms their equivalence (Tables 1 to 3, Fig. 1).

A Simple Correspondence Analysis is conducted on both the original contingency table, with values ranging from 0 to 1, and the scaled table, which adjusts these values to a 0–100 range. It is expected that, although the numerical values in the scaled table cells are larger, the relative proportions among the cells remain unchanged.

To implement the factor analysis strategy for correspondences and association tests, the factor analysis module of the commercial software XLSTAT is employed [16]. The resulting maps from each analysis—both for the original and the scaled tables—should exhibit similar clustering patterns among the corresponding categories. This indicates that the transformations applied (i.e., linear scaling) have not altered the underlying relationships between the variable categories.

Table 1: On the left, original example table; on the right, table scaled from 0 to 100.

Values of 0 to 1

Ori	C1	C2	C3	C4
F1	0	0	0	1
F2	0	0	0	1
F3	0	1	0	0
F4	0	1	0	0
F5	0	0	1	0
F6	0	0	0	1
F7	1	0	0	0

Scalation of 0 to 100

Scale	C1	C2	C3	C4
F1	0	0	0	100
F2	0	0	0	100
F3	0	100	0	0
F4	0	100	0	0
F5	0	0	100	0
F6	0	0	0	100
F7	100	0	0	0

Table 2: Inertia per cell.

ORIGINAL

Inertia per cell:

	C1	C2	C3	C4
F1	0,02041	0,04082	0,02041	0,10884
F2	0,02041	0,04082	0,02041	0,10884
F3	0,02041	0,25510	0,02041	0,06122
F4	0,02041	0,25510	0,02041	0,06122
F5	0,02041	0,04082	0,73469	0,06122
F6	0,02041	0,04082	0,02041	0,10884
F7	0,73469	0,04082	0,02041	0,06122

Eigen values and percentages of inertia:

	F1	F2	F3
Eigen value	1,000	1,000	1,000
Inertia (%)	33,333	33,333	33,333
% accumulated	33,333	66,667	100,000

Table 3: Inertia per cell.

ESCALATION 0 to 100

Inertia per cell:

	C1	C2	C3	C4
F1	0,02041	0,04082	0,02041	0,10884
F2	0,02041	0,04082	0,02041	0,10884
F3	0,02041	0,25510	0,02041	0,06122
F4	0,02041	0,25510	0,02041	0,06122
F5	0,02041	0,04082	0,73469	0,06122
F6	0,02041	0,04082	0,02041	0,10884
F7	0,73469	0,04082	0,02041	0,06122

Eigen values and percentages of inertia:

	F1	F2	F3
Eigen value	1,000	1,000	1,000
Inertia (%)	33,333	33,333	33,333
% accumulated	33,333	66,667	100,000

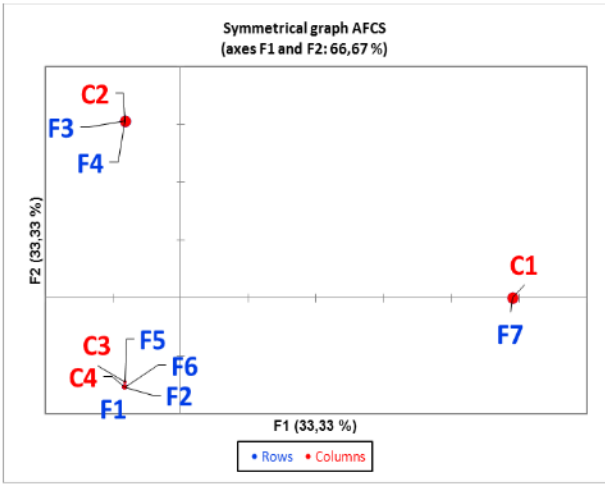

Fig. 1: Resulting maps from each analysis.

A key element is the total inertia explained by the first two axes. If the structure of the information remains consistent, the inertia (which reflects the variance explained by these axes) should be equivalent between the two tables, adjusted for the effect of scaling. It is important to note that the significant dimensions (axes) emerging from the correspondence analysis, as well as the contributions of the categories to these axes, are consistent between the original and the scaled tables. The relative positions of the categories within the correspondence space should remain stable, indicating that the structural relationships are preserved despite the scaling (Tables 4 to 7 and Tables 8 to 11).

Table 4: Test of independence using Chi-Square.

ORIGINAL

Test of independence between rows and columns (Chi-square):

Chi-square (Observed value)	21
Chi-square (Critical value)	28,869
GL	18
p-value	0,279
alfa	0,05

Test interpretation:

H0: The rows and columns of the table are independent.

Ha: There is a dependency between the rows and columns of the table.

Since the calculated p-value is greater than the significance level alpha=0.05, the null hypothesis H0 cannot be rejected.

Theoretical frequencies less than 5 have been detected. To safely use the Chi-square test based on the approximation by the Chi-square distribution, the theoretical frequencies should not be less than 5.

Table 5: Test of independence using Wilks' G^2.

Test for independence between rows and columns (Wilks' G^2):

Wilks G^2 (Observed Value)	17,878
Wilks G^2 (Critical Value)	28,869
GL	18
p-value	0,464
alfa	0,05

Test interpretation:

H0: The rows and columns of the table are independent.

Ha: There is a dependency between the rows and columns of the table.

Since the calculated p-value is greater than the significance level alpha=0.05, the null hypothesis H0 cannot be rejected.

Table 6: Association coefficients.

Association coefficients:

Coefficient	Value
Pearson Phi	1,732
Contingency coefficient	0,866
Cramer's V	1,000
T for Tschuprow	0,841
Tau by Goodman and Kruskal (F/C)	0,500
Goodman and Kruskal Tau (C/F)	1,000

Table 7: Significance per cell using Fisher's Exact Test.

Significance per cell (Fisher's exact test):

	C1	C2	C3	C4
F1	<	<	<	>
F2	<	<	<	>
F3	<	>	<	<
F4	<	>	<	<
F5	<	<	>	<
F6	<	<	<	>
F7	>	<	<	<

Values in red are significant at the alpha=0.05 level.

Table 8: Test of independence using Chi-Square.

SCALATION 0 a 100

Test of independence between rows and columns (Chi-square):	
Chi-square (Observed value)	2100
Chi-square (Critical value)	28,869
GL	18
p-value	< 0,0001
alfa	0,05

Test interpretation:
H0: The rows and columns of the table are independent.
Ha: There is a dependency between the rows and columns of the table.
Since the computed p-value is less than the significance level alpha=0.05, the null hypothesis H0 should be rejected, and the alternative hypothesis Ha should be accepted.

Table 9: Test of independence using Wilks' G^2.

Test for independence between rows and columns (Wilks' G^2):	
Wilks G^2 (Observed Value)	1787,848
Wilks G^2 (Critical Value)	28,869
GL	18
valor-p	< 0,0001
alfa	0,05

Test interpretation:
H0: The rows and columns of the table are independent.
Ha: There is a dependency between the rows and columns of the table.
Since the computed p-value is less than the significance level alpha=0.05, the null hypothesis H0 should be rejected, and the alternative hypothesis Ha should be accepted.

Table 10: Association coefficients.

Association coefficients:	
Coeficiente	Value
Pearson Phi	1,732
Contingency coefficient	0,866
Cramer's V	1,000
T for Tschuprow	0,841
Tau by Goodman and Kruskal (F/C	0,500
Goodman and Kruskal Tau (C/F)	1,000

Table 11: Significance per cell using Fisher's Exact Test.

Significance per cell (Fisher's exact test):

	C1	C2	C3	C4
F1	<	<	<	>
F2	<	<	<	>
F3	<	>	<	<
F4	<	>	<	<
F5	<	<	>	<
F6	<	<	<	>
F7	>	<	<	<

Values in red are significant at the alpha=0.05 level.

The application of the proposed methodology preserves the proportions of the original table while scaling the values to enable statistical testing.

- It allows for the analysis of tables with small values, where standard statistical tests may not be valid.
- Scaling to a broader range increases the statistical power to detect associations.
- The linear transformation preserves the original relationships among cells, ensuring that the results of tests such as Chi-Square remain unchanged.
- It facilitates the interpretation of table values expressed on a percentile scale.
- It enables simple correspondence analysis with greater statistical validity.

It is inferred that the proposed linear scaling allows for the full utilization of information in tables with small values, rendering the analysis more robust and powerful without altering the original relationships. Its application in the analysis of fuzzy data and scenarios with low expected frequencies is highly promising.

3 A MORE COMPLEX EXAMPLE

In [17], Bolboaca and co-authors employed Fisher's Exact Test to illustrate the analysis of independence in the experimental data reported by Fisher [18] (Table 12). In this study, Fisher's table is used to provide a more complex exemplification of the proposed methodology. These experimental values correspond to the response to manure fertilization across different potato varieties. The comparison is made between Treatment and Variety; UD, KK, KP, TP, ID, GS, AJ, BQ, ND, EP, AC, DY: potato varieties (UD = Up to Date; KK = King of K; KP = Kerr's Pink; TP = Tinwald Perfection; ID = Iron Duke; GS = Great Scott; AJ = Ajax; DY = Duke of York); DS, DC, US, UC, UB: treatment types (D* – manure; U* – no manure; S – sulfate; C – chloride; B – basal). Based on this Fisher table, established methods were applied to compare the original table and the scaled version (Tables 12 to 16 and Tables 17 to 21, respectively).

Table 12: Experimental values: response to manure fertilization across different potato varieties (original Fisher table).

ORIGINAL

TV	UD	KK	KP	TP	ID	GS	AJ	BQ	ND	EP	AC	DY
DS	25	28	23	20	23	21	22	22	18	15	14	10
DC	26	27	24	19	21	24	17	21	20	16	11	12
DB	27	24	14	20	20	22	22	21	16	14	11	13
US	23	20	18	20	16	16	13	13	12	13	13	8,2
UC	19	17	21	18	18	14	20	14	13	12	13	8,3
UB	9,5	6,5	4,9	7,7	4,4	2,3	4,2	6,6	1,6	2,2	2,2	1,6

Table 13: Test of independence using Chi-Square.

ORIGINAL

Test of independence between rows and columns (Chi-square):

Chi-square (Observed value)	22,022
Chi-square (Critical value)	73,311
GL	55
p-value	1,000
alfa	0,05

Test interpretation:

H0: The rows and columns of the table are independent.

Ha: There is a dependency between the rows and columns of the table.

Since the calculated p-value is greater than the significance level alpha=0.05, the null hypothesis H0 cannot be rejected.

Theoretical frequencies less than 5 have been detected. To safely use the Chi-square test based on the approximation by the Chi-square distribution, the theoretical frequencies should not be less than 5.

Table 14: Test of independence using Wilks' G^2.

Test for independence between rows and columns (Wilks' G^2):

Wilks G^2 (Observed Value)	22,505
Wilks G^2 (Critical Value)	73,311
GL	55
p-value	1,000
alfa	0,05

Test interpretation:

H0: The rows and columns of the table are independent.

Ha: There is a dependency between the rows and columns of the table.

Since the calculated p-value is greater than the significance level alpha=0.05, the null hypothesis H0 cannot be rejected.

Table 15: Association coefficients.

Association coefficients:

Coefficient	Value
Pearson Phi	0,14
Cramer's V	0,06

Table 16: Significance per cell using Fisher's Exact Test.

Significance per cell (Fisher's exact test):

	UD	KK	KP	TP	ID	GS	AJ	BQ	ND	EP	AC	DY
DS	<	>	>	<	>	<	>	>	>	<	>	<
DC	<	>	>	<	<	>	<	>	>	>	<	>
DB	>	<	<	<	<	>	>	>	<	>	<	>
US	>	>	>	>	<	<	<	<	<	>	>	<
UC	<	<	>	>	>	<	>	<	<	>	>	<
UB	>	>	<	>	<	<	<	>	<	<	<	<

Values in red are significant at the alpha=0.05 level.

Table 17: Original Fisher table scaled from 1 to 100.

SCALATION 0 to 100

TV	UD	KK	KP	TP	ID	GS	AJ	BQ	ND	EP	AC	DY
DS	90	100	82	70	81	73	78	77	63	50	46	32
DC	92	96	86	66	72	86	58	73	71	53	36	39
DB	94	84	48	70	70	77	76	72	55	48	36	44
US	81	71	63	70	54	54	42	42	39	41	41	25
UC	64	58	73	63	60	48	68	46	43	39	42	25
UB	30	19	13	23	11	3	10	19	0	2	2	0

Table 18: Test of independence using Chi-Square.

SCALATION 0 to 100

Test of independence between rows and columns (Chi-square):

Chi-square (Observed value)	118,334
Chi-square (Critical value)	73,311
GL	55
p-value	< 0,0001
alfa	0,05

Test interpretation:

H0: The rows and columns of the table are independent.

Ha: There is a dependency between the rows and columns of the table.

Since the computed p-value is less than the significance level alpha=0.05, the null hypothesis H0 should be rejected, and the alternative hypothesis Ha should be accepted.

Table 19: Test of independence using Wilks' G^2.

Test for independence between rows and columns (Wilks' G^2):

Wilks G^2 (Observed Value)	134,224
Wilks G^2 (Critical Value)	73,311
GL	55
p-value	< 0,0001
alfa	0,05

Test interpretation:

H0: The rows and columns of the table are independent.

Ha: There is a dependency between the rows and columns of the table.

Since the computed p-value is less than the significance level alpha=0.05, the null hypothesis H0 should be rejected, and the alternative hypothesis Ha should be accepted.

Table 20: Association coefficients.

Association coefficients:

Coefficient	Value
Pearson Phi	0,18
Cramer's V	0,08

Table 21: Significance per cell using Fisher's Exact Test.

Significance per cell (Fisher's exact test):

	UD	KK	KP	TP	ID	GS	AJ	BQ	ND	EP	AC	DY
DS	<	>	>	<	>	<	>	>	>	<	>	<
DC	<	>	>	<	<	>	<	>	>	>	<	>
DB	>	<	<	<	<	>	>	>	<	>	<	>
US	>	>	>	>	<	<	<	<	<	>	>	<
UC	<	<	>	>	>	<	>	<	<	>	>	<
UB	>	>	>	>	<	<	<	>	<	<	<	<

Values in red are significant at the alpha=0.05 level.

The benefits of applying linear scaling to adjust the values in contingency tables for statistical analysis are as follows:

- Homogenization of the value scale: Transforming all data to a common scale from 0 to 100 minimizes the impact of the wide variability in original ranges. This normalization not only facilitates direct comparisons between groups but also ensures a more uniform and equitable application of statistical tests.
- Increased statistical power: Expanded scaling significantly enhances the ability to detect associations in cells containing small values and to identify subtle yet statistically significant patterns. This modification is essential for revealing trends that might otherwise remain undetected under more restrictive scales.
- Compliance with statistical assumptions: Adjusting the values increases those that are small, thereby facilitating compliance with critical assumptions for tests such as the Chi-Square, which require expected frequencies greater than five in each cell.
- Ease of interpretation: Employing a common metric from 0 to 100 simplifies the visual interpretation of data, making trends and patterns more evident—patterns that might be difficult to discern on the original scale.
- Applicability of multiple statistical tests: Standardization through scaling enables the application of a broader range of statistical tests that require specific conditions in the data, thereby maximizing the utility of the contingency table.
- Preservation of association structure: Despite the transformation of values, the row/column association structure of the original table remains intact, ensuring that the conclusions drawn are consistent with the underlying data structure.
- Comprehensive view of relationships: By complementing traditional tests with scaling, a more robust and comprehensive perspective of the relationships between variables is achieved, enhancing the overall interpretation of the data.
- Although scaling may introduce variations in the inertia of contingency tables, it does not compromise the original data structure. In fact, this adjustment allows for a more detailed evaluation of underlying associations, offering a richer and more nuanced view that may be critical in research contexts where small values predominate.

In conclusion, the application of linear scaling balances analytical conditions and enhances analytical capacity without distorting the intrinsic relationships within the data. As a result, this technique facilitates the identification of findings that might otherwise remain invisible without such adjustment, thereby improving the quality and precision of statistical analysis.

4 CASE OF TABLES WITH FUZZY NUMBERS

In the case of tables containing fuzzy numbers or possibility values, once equivalence is established through linear scaling, a simple correspondence analysis may be performed. By preserving the original proportions of the table, linear scaling maintains the fuzzy relationships between cells. For instance, if cell A has twice the possibility value of cell B in the original table, this ratio is preserved in the scaled version.

This aspect is essential when working with fuzzy logic, as the relative distribution of possibility values conveys critical information regarding the fuzzy nature of the categories. Scaling the table does not distort these relationships, thereby allowing the same interpretative conclusions to be drawn from the fuzzy data. Moreover, expanding the values to a broader range through scaling enhances the resolution of intermediate possibilities. For example, a probability of 0.2 may be expanded to 20, offering clearer differentiation from a value of 0.3, which would be scaled to 30.

This provides greater degrees of freedom for statistical tests, rendering them more robust and precise in detecting diffuse relationships and associations within the data. Tests such as Chi-Square or Fisher's Exact Test tend to be less reliable when expected frequencies are very low.

Finally, in fuzzy tables, possibility values are typically low, often concentrated between 0 and 0.5. By scaling them to a broader range, the statistical power of the tests is increased, enabling the identification of diffuse relationships that might otherwise go unnoticed due to the small magnitude of the values.

5 COMPARATIVE ADVANTAGES USING SCALED CONTINGENCY TABLES

The linear scaling of contingency tables—particularly those incorporating fuzzy numbers or possibility values—is performed to adapt the data structure to the requirements of more robust

statistical analyses, such as Simple Correspondence Analysis (SCA). This transformation process is crucial when the original tables contain minimum values that do not correspond to zero, which could distort data interpretation if not properly adjusted.

5.1 Effects of Linear Scaling

Preservation of proportional structure: When applying a linear scale, the original proportional structure of the table is maintained. For example, if cell A in the original table contains a possibility value twice that of cell B, this ratio is preserved after scaling (e.g., 0.2 to 20 and 0.1 to 10). This consistency is essential to ensure that the interpretations derived from the analysis are not affected by the transformation of the data.

Increased resolution of intermediate possibilities: By expanding values to a broader numerical range, scaling improves the resolution of the data. Subtle differences between possibility values become more distinguishable, increasing the sensitivity of subsequent statistical tests.

This is particularly valuable in contexts where small differences may carry significant practical implications.

5.2 Impact on Simple Correspondence Analysis

Variability in inertias: Although scaling preserves the proportional structure, the inertias calculated in SCA may vary. This is because inertia depends on the absolute distribution of frequencies within the table, and scaling modifies these absolute values. However, since the relative pattern among cells remains constant, the qualitative interpretation of the correspondence maps should not change significantly.

Robustness in detecting fuzzy relationships: Given that possibility values are typically low—often concentrated between 0 and 0.5—scaling broadens the analytical spectrum by converting them into a wider range, such as 0 to 100. This adjustment not only enhances the ability to detect subtle and diffuse relationships among categories, but also increases the statistical power of tests like Chi-Square and Fisher's Exact Test, which are less effective when expected frequencies are low.

6 VAGUE CONTINGENCY DATA

Aslam and Alamri [19] describe a method that modifies Fisher's Exact Test using neutrosophic statistics, focusing on the analysis of contingency data that are not precisely defined. This approach enables the use of neutrosophic numbers, which represent data with inherent uncertainty through a format that incorporates degrees of truth, falsehood, and indeterminacy for each data element.

This method differs from the traditional Fisher's Exact Test, which employs precise and well-defined values to calculate the probability of observing a data distribution at least as extreme as the one observed, under the null hypothesis that no differences exist between groups. In contrast, the neutrosophic modification allows uncertainty and indeterminacy to be incorporated directly into the data, which is particularly useful in contexts such as social surveys or perception studies, where responses are not always clear or direct. In practice, the principles of neutrosophic theory are applied to expand the contingency matrices used in Fisher's test with additional dimensions that capture these degrees of indeterminacy, thereby enabling a richer and more realistic interpretation of the data under analysis.

The approach proposed in this study addresses uncertainty in the data by applying a one-time preprocessing step prior to conducting classical statistical tests such as Fisher's Exact Test.

The data are transformed into fuzzy sets before statistical analysis, maintaining the use of traditional statistical tests, albeit adapted to the new data structure.

- Identification of Imprecision: Imprecision in the data is identified as a consequence of various factors, including missing values, measurement error, subjectivity, and the complexity of the phenomenon under study. The uncertainty inherent in this context requires specialized handling, for which the use of fuzzy logic tools is essential.
- Use of fuzzy variables: The original variables are transformed into fuzzy sets, representing imprecision through degrees of membership rather than binary or discrete values. This allows for a more flexible and adaptive interpretation of the data, avoiding the rigidity of strict and potentially inaccurate categorizations.
- Construction of Contingency Tables with Possibility Values: Contingency tables are constructed incorporating these degrees of membership, with values scaled to a range from 0 to 100. This is a fundamental step in adapting Fisher's Exact Test and other association tests to the diffuse nature of the data.
- Application of Statistical Tests: Despite the transformation of the data into fuzzy form, conventional statistical methods such as Fisher's Exact Test are used to evaluate

independence or association between variables, with these tests adapted to work effectively with the new data structures.

6.1 Comparison of Both Methods

Both methods aim to address the challenge of working with data that contain uncertainty or vagueness, albeit through distinct theoretical frameworks and data transformation strategies. While our approach in [18] employs fuzzy set theory to modify the data prior to applying established statistical tests, the neutrosophic method adapts the statistical test itself to directly accommodate uncertain data without prior transformation at the data value level.

- Regarding the use of fuzzy sets: The proposed method involves transforming the original variables into fuzzy sets, which entails defining degrees of membership of the data to specific sets, based on the membership function that best characterizes the uncertainty or vagueness of the data. For instance, rather than relying on classical binary data, an element may belong to a set with a degree of 0.7, indicating a level of membership rather than a strict binary classification. In contrast, the neutrosophic method utilizes neutrosophic numbers that incorporate degrees of truth, falsehood, and indeterminacy. Each element in the analysis provides information about its potential truthfulness, falsity, and a degree of uncertainty that is typically not considered in conventional statistical approaches.
- Regarding the construction of contingency tables: The proposed method constructs contingency tables using possibility values derived from fuzzy sets, which reflect the uncertainty in the classification of each entry. These values are scaled to a range of 0 to 100 to standardize and facilitate the application of association tests. This scaling helps to manage small values that could compromise the validity of traditional statistical tests. In the neutrosophic method, contingency tables are modified to incorporate indeterminacy alongside truth and falsehood, thereby expanding the classical contingency matrix into a more complex structure that better captures the uncertainty inherent in the data.
- Regarding the application of statistical tests: Once the contingency table has been adjusted to reflect uncertainty through fuzzy sets and appropriately scaled, Fisher's Exact Test is applied to assess independence or association between variables. This step maintains a robust statistical approach, adapted to the new data structure. In the neutrosophic method, the application of neutrosophic statistics may require modifying the calculation of Fisher's Exact Test to incorporate degrees of indeterminacy, potentially by adjusting the computation of the exact probability of the observed tables under the null hypothesis.

The methodology proposed in this study offers interpretative flexibility and is compatible with existing theories that utilize multiple ranges or categories, thereby enabling a more nuanced understanding. Both approaches provide valuable frameworks for managing uncertainty in statistical analysis, each pursuing different theoretical and practical paths to integrate or accommodate imprecision in data. The approach presented herein is distinguished by its adaptability and its capacity to incorporate fuzzy logic in a manner that is more intuitive and closely aligned with the way humans interpret imprecise information.

7 DATA WITH SMALL SAMPLE SIZES

In [20], the authors present an independent contingency testing approach using the bootstrap method to enhance precision in contingency tables, particularly in cases involving small sample sizes where Chi-Square and Fisher's Exact Test may be less effective. The Chi-Square and Fisher tests are discussed and critiqued for being asymptotic and conservative, respectively, which can lead to errors, especially in tables with low cell frequencies. To address this issue, bootstrap versions of the Chi-Square tests (both Pearson and likelihood ratio) are proposed. These simulate the distribution of the test statistic under the null hypothesis through resampling, resulting in more accurate tests that are less dependent on sample size. Simulation studies are used to demonstrate that bootstrap tests maintain the nominal level more accurately than traditional asymptotic approximations and Fisher's Exact Test.

Both methods aim to improve the validity of independence tests in contingency tables under conditions of uncertainty or small sample sizes, but they do so from different operational philosophies: modifying the data versus modifying the testing process.

The method presented in this study stands out for its ability to integrate fuzzy logic in a way that is more accessible and relevant to those directly involved in

the collection or interpretation of ambiguous data.

- Theoretical foundation: This study proposes the application of fuzzy logic transformations to the variables prior to analysis, creating contingency tables based on degrees of membership, which can better reflect the uncertainty present in the data. The bootstrap method seeks to improve the precision of standard tests through resampling, without altering the nature of the data, but rather adjusting the evaluation of the test statistic.
- Data transformation: A data structure is proposed that results from converting values into fuzzy degrees of membership prior to any statistical analysis. The bootstrap method does not require modification of the data; instead, it adapts the statistical evaluation process to be more robust in the context of small samples and low frequencies.
- Evaluation and comparison: An intuitive approach is offered, directly aligned with human interpretations of imprecise data, adapting existing tests for use with new data formats. It is proposed to scale contingency tables incorporating these degrees of membership from a 0–1 range to a 0–100 range. This scaling allows for the adaptation of traditional statistical tests such as Fisher's Exact Test to conditions where small or near-zero values could negatively affect the test's validity. The bootstrap method provides a robust and flexible approach that can be applied without prior modifications to the data, which may be advantageous in applications requiring the preservation of the original data integrity or when the optimal way to parameterize uncertainty is unknown.
- Adaptation of the testing process: Similar to the bootstrap method, the approach presented in this thesis also adapts the testing process to better handle suboptimal sampling conditions. However, this is achieved by modifying the data structure to reflect uncertainty, unlike bootstrapping, which adjusts the statistical evaluation procedure.

8 CONCLUSIONS

The linear scaling approach is ideal for fully leveraging the information contained in fuzzy tables, rendering the analysis more robust and sensitive to uncovering complex underlying fuzzy relationships. Although the scaling process modifies the absolute values of the entries in the contingency table, it does not affect the conclusions that can be drawn from Simple Correspondence Analysis. This validates the use of scaling as a technique that facilitates the application of statistical tests without compromising the analytical integrity of the data.

Correspondence analysis applied to scaled tables demonstrates that, while the relative proportions and relationships between cells are preserved, the analysis benefits from higher resolution and an expanded range, allowing for a deeper exploration of fuzzy associations. These findings confirm the effectiveness of linear scaling as a technique for preparing fuzzy data for complex statistical analyses, while ensuring that the fundamental interpretation of the data remains intact.

Linear scaling not only enables the robust application of tests such as Chi-Square and Fisher's Exact Test to tables with very low frequencies, but also provides a practical tool for researchers in diverse fields such as agronomy, medicine, engineering, and sensor analytics. The methodology developed, along with its associated software, allows for the automatic transformation of data from a 0–1 scale to a 0–100 scale, preserving cell relationships and enhancing the detection of patterns that might otherwise remain hidden.

This opens new avenues for the analysis of imprecise data—such as membership matrices in recommender systems or ultrafine measurements in geoscience and biometrics. It is recommended to integrate this approach with other rescaling techniques and to extend its application to multidimensional tables, time series, and machine learning algorithms. Likewise, the development of software packages in various programming languages and the comparison of this method's performance with alternative approaches will contribute to its consolidation within the scientific community.

This study represents a significant advancement in the management of imprecise data and offers new pathways for exploring categorized data across various scientific disciplines and industrial applications. For future research, it would be beneficial to explore the application of linear scaling to other types of statistical data that face similar challenges, such as biometric or geostatistical data, where measurements are often extremely precise but small in magnitude.

REFERENCES

[1] Y. Zhai, W. Song, X. Liu, L. Liu, and X. Zhao, "A Chi-Square Statistics Based Feature Selection Method in Text Classification," in *IEEE 9th International Conference on Software Engineering*

and Service Science (ICSESS), Beijing, China, 1918.
[2] S. Rosidin, G. Fajar Shidik, and A. Zainul Fanani, "Improvement with Chi Square Selection Feature using Supervised Machine Learning Approach on Covid-19 Data," *International Seminar on Application for Technology of Information and Communication*, Semarangin, Indonesia, 2021.
[3] J. Angulo-Paniagua, and J. Quirós-Tortós, "Comparing Chi-square-Based Bad Data Detection Algorithms for Distribution System State Estimation," *IEEE PES Transmission & Distribution Conference and Exhibition - Latin America (T&D LA)*, Montevideo, Uruguay, 2020.
[4] Z Wang, Z. Huang, Y. Xu, Y. Zhang, and X. Li, "Image Noise Level Estimation by Employing Chi-Square Distribution," *21st International Conference on Communication Tech*, Tianjin, China, 2021.
[5] I. Chen, "A Novel and Fast Distributed Computation Method for Fisher's Exact Test and Its Application in Gene Expression Profiling Studies," *IEEE MIT Undergraduate Research Technology Conference (URTC)*, Cambridge, MA, USA, 2022.
[6] A. Poon, S. Jankly, and T. Chen, "Privacy Preserving Fisher's Exact Test on Genomic Data," *IEEE International Conference on Big Data (Big Data)*, Seattle, WA, USA, 2018.
[7] P. Campos, I. Cantador, F. Díez, and I. Fernández-Tobías, "A Criterion Based on Fisher's Exact Test for Item Splitting in Context-Aware Recommender Systems," *33rd International Conference of the Chilean Computer Science Society (SCCC)*, Talca, Chile, 2014.
[8] R. Yano, H. Tanioka, K. Matsuura, M. Sano, and T. Ueta, "Quantitative Measurement and Analysis to Thinking as a Way of Programming for Elementary School in Japan," *9th International Congress on Advanced Applied Informatics (IIAI-AAI)*, Kitakyushu, Japan, 2020.
[9] H. Kim, "Statistical notes for clinical researchers: Chi-squared test and Fisher's exact test," *Restorative Dentistry & Endodontics*, vol. 42, pp. 152-155, 2017.
[10] K.T. Jafseer, S. Shailesh, and A.Sreekumar, "Feature Drift Detection using Overlapping Window and Mann-Whitney U Test," *4th International Conference on Innovative Trends in Information Technology (ICITIIT)*, Kottayam, India, 2023.
[11] M. Fetaji, L. Morina, and B. Fetaji, "Devising and evaluating B2B conceptual model for B2B portal for mobile interactive devices using Mann-Whitney U test," *6th Mediterranean Conference on Embedded Computing (MECO)*, Bar, Montenegro, 2017.
[12] K. Jafseer, S. Shailesh, and A. Sreekumar, "Feature Drift Detection using Overlapping Window and Mann-Whitney U Test," *4th International Conference on Innovative Trends in Information Technology (ICITIIT)*, Kottayam, India, 2023.
[13] M. Césari, N. Ventrera, and A. Gámbaro, "Análisis de datos sensoriales de tomate triturado con lógica difusa y técnicas multivariadas.", *Revista de la Facultad de Ciencias Agrarias, Universidad Nacional de Cuyo*, vol. 50(1), pp. 233-248, 2028.
[14] M. Césari, N. Ventrera, and A. Gámbaro, "Análisis de datos sensoriales de tomate triturado con lógica difusa y técnicas multivariadas," *Revista de la Facultad de Ciencias Agrarias, Universidad Nacional de Cuyo*, vol. 50(1), pp. 233-248, 2018.
[15] M. Césari, and R. Césari, "La lógica difusa aplicada para valorar rúbricas de evaluación," *X Congreso Nacional de Ingeniería Informática / Sistemas de Información - CoNaIISI 2022*, Facultad
[16] Addinsoft Xlstat versión 2018, licencia para investigación. Análisis estadístico para Microsoft Excel desarrollada por Addinsoft 1996-2018, Available in www.xlstat.com/es/products/xlstat-pro/.
[17] S. Bolboacă, L. Jäntschi, A. Sestraş, R. Sestraş, and D. Pamfil, "Fisher chi-square statistic revisited. Information," *Pearson*, vol. 2(3), pp. 528-545, 2011.
[18] R. Fisher, "The conditions under which χ 2 measures the discrepancy between observation and hypothesis," *Journal of the Royal Statistical Society,* pp. 442-450, 1924.
[19] M. Aslam, and F. Alamri, "Data analysis for vague contingency data.", *Journal of Big Data*, vol. 10(1), pp. 131, 2023.
[20] J. Lin, C. Chang, and N. Pal, "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*, vol. 25(3), pp. 438-458, 2015, DOI: 10.1080/10543406.2014.920851.
[21] Césari, M. (2024). *DaFu: Automated Linear Scaling for Statistical Analysis in Contingency Tables* [Computer software]. GitHub. https://github.com/matucesari/DaFu
[22] Césari, M. (2024). *DaFu03: Fuzzy Associations Web Application* http://micesari.servehttp.com:3838/DaFu03_AsociacionesDifusas/