

International Journal of Statistics and Applied Mathematics

ISSN: 2456-1452
Maths 2025; 10(6): 292-295
© 2025 Stats & Maths
<https://www.mathsjournal.com>
Received: 06-04-2025
Accepted: 10-05-2025

Jakub Zielinski
Department of Algebra and
Geometry, Wrocław Academy of
Statistical Research, Wrocław,
Lower Silesia, Poland

Anna Szymanska
Department of Algebra and
Geometry, Wrocław Academy of
Statistical Research, Wrocław,
Lower Silesia, Poland

Michał Wojciechowski
Department of Algebra and
Geometry, Wrocław Academy of
Statistical Research, Wrocław,
Lower Silesia, Poland

Corresponding Author:
Jakub Zielinski
Department of Algebra and
Geometry, Wrocław Academy of
Statistical Research, Wrocław,
Lower Silesia, Poland

Resolving contingency table analysis with structural zeros with applications to ordos plateau energy resource mapping

Jakub Zielinski, Anna Szymanska and Michał Wojciechowski

DOI: <https://www.doi.org/10.22271/math.2025.v10.i6d.2483>

Abstract

In the early twentieth century, Karl Pearson's chi-square test assumed that every cell in a contingency table could, in principle, contain an observation—but energy resource surveys routinely produce tables where certain category combinations are structurally impossible. This research develops a modified contingency-table framework that handles structural zeros by restricting the chi-square summation to feasible cells and adjusting the degrees of freedom accordingly. The method was tested on a 6×8 cross-classification of geological formation type and energy resource category across 214 survey sites on the Ordos Plateau, collected between June 2020 and November 2023. Of the 48 cells, 11 were structural zeros—combinations such as 'unconsolidated sand' $\times$ 'geothermal' that cannot co-occur geologically. The corrected chi-square statistic was 87.34 with 26 adjusted degrees of freedom ($p < 0.001$), whereas the standard test yielded an inflated value of 143.71 with 35 degrees of freedom, over-rejecting the null by 64.5%. Simulation with 10,000 Monte Carlo tables confirmed that the corrected test maintained the nominal 5% Type I error rate (observed: 4.83%), while the uncorrected version exceeded 12%. Cramér's V, recalculated over feasible cells only, was 0.418, indicating a moderate-to-strong association between formation type and resource presence. The results argue for routine adoption of structural-zero corrections in geoscience contingency analyses.

Keywords: Contingency table, structural zeros, chi-square test, degrees of freedom, Ordos Plateau, energy resources, geological formation, Cramér's V, Monte Carlo simulation, categorical data analysis

Introduction

When Karl Pearson proposed the chi-square goodness-of-fit test in 1900, he worked with dice rolls and card games where every outcome was physically possible ^[1]. Modern contingency tables, especially those arising in natural-resource surveys, often contain cells that are empty not by sampling accident but by structural impossibility—certain geological formations simply cannot host certain resource types ^[2]. Treating these structural zeros as ordinary zero counts inflates both the test statistic and the degrees of freedom, leading to distorted p-values and unreliable association measures ^[3].

The Ordos Plateau in north-central China is one of Asia's most intensively surveyed energy basins, holding coal, natural gas, coalbed methane, and expanding solar installations ^[4]. Geological mapping has identified at least six major formation classes, from Carboniferous limestone to Quaternary loess, each with a distinct resource portfolio ^[5]. Cross-tabulating formation type against resource category is a natural first step in spatial resource mapping, yet the inevitable structural zeros have forced analysts to rely on ad hoc adjustments or to collapse table margins until the zeros disappear ^[6]. Neither approach is satisfying: the first lacks a formal statistical basis, and the second discards information ^[3].

Bishop, Fienberg, and Holland ^[7] laid the groundwork for incomplete contingency tables in the 1970s, and Haberman ^[8] developed the asymptotic theory for log-linear models with fixed zeros. But practical guidance for geoscientists remains sparse. This research bridges that gap by deriving the adjusted chi-square test explicitly, calibrating it through Monte Carlo simulation, and demonstrating its use on a real Ordos Plateau dataset ^[9].

The three objectives were: (a) to formulate the restricted chi-square statistic and its degrees-of-freedom correction, (b) to validate Type I error control via simulation, and (c) to compare corrected and uncorrected association measures for resource mapping ^[10]. Analytical work was conducted at the Wroclaw Academy of Statistical Research, Wroclaw, Poland, using geological data shared under agreement WAG-2020-018.

System Design

The analytical pipeline consisted of three modules. The first module ingested raw survey records (site coordinates, formation class, resource presence/absence) and assembled the 6×8 contingency table, flagging structural-zero cells from a geological compatibility matrix provided by domain experts ^[5]. The second module computed the corrected chi-square statistic by summing $(O - E)^2/E$ only over the f feasible cells, where expected values E were obtained by iterative proportional fitting restricted to the feasible set ^[8]. The degrees of freedom were set to $f - r - c + 1 + z$, with z being the number of structural zeros, r the row count, and c the column count ^[3]. The third module ran Monte Carlo validation by generating 10,000 tables under the null hypothesis of independence (conditional on feasible cells), recording the empirical distribution of the corrected statistic ^[9].

Materials and Methods

Materials

Geological survey data were obtained from the Ordos Plateau Regional Geological Bureau, covering 214 sites surveyed between June 2020 and November 2023. Each site was classified into one of six formation types: Carboniferous limestone, Permian sandstone, Triassic mudstone, Jurassic coal-bearing strata, Cretaceous red beds, and Quaternary loess ^[4]. Resource categories were: coal, natural gas, coalbed methane, oil shale, geothermal, solar suitability, wind suitability, and no detectable resource ^[5]. The geological compatibility matrix, which identified the 11 structural-zero cells, was constructed by a panel of three senior geologists and cross-checked against published stratigraphic charts ^[6]. All statistical analyses were performed at the Wroclaw Academy of Statistical Research under project approval ALG-2020-044.

Methods: Expected cell frequencies under the null hypothesis of independence were computed via iterative proportional fitting (IPF), restricting the iteration to the 37 feasible cells ^[8]. Convergence was declared when the maximum absolute difference between successive expected-value vectors fell below 10^{-8} . The corrected chi-square statistic χ^2_c was then calculated as $\sum_{\{feasible\}} (O_{ij} - E_{ij})^2 / E_{ij}$ with 26 degrees of freedom ^[3]. For comparison, the standard Pearson χ^2 was computed over all 48 cells with 35 degrees of freedom. Cramér's V was calculated in both the corrected (V_c) and uncorrected (V_u) forms, using the minimum of $(r - 1, c - 1)$ after adjustment for structural zeros in the corrected version ^[7]. Monte Carlo validation sampled 10,000 tables from a multinomial distribution with cell probabilities proportional to the IPF-fitted expected values and structural-zero cells fixed at zero ^[9]. All computations were implemented in R 4.3.2 with the 'MASS' and 'vcdExtra' packages on a 3.40 GHz Intel i7 workstation with 32 GB RAM.

Implementation Details: The IPF algorithm was coded as a double loop over row and column marginals, with structural-zero cells masked via a binary indicator matrix Z ($z_{ij} = 0$ for structural zeros, 1 otherwise) ^[8]. Each IPF iteration required two sweeps: row-normalisation followed by column-normalisation, both restricted to cells where $z_{ij} = 1$. Convergence typically occurred within 15-20 iterations. The Monte Carlo sampling used R's 'rmultinom' function seeded at 42 for reproducibility. Parallel execution across four CPU cores (via the 'parallel' package) reduced the 10,000-replication loop from 47 seconds to 13 seconds ^[9].

Results

Table 1: Chi-square test results: corrected versus standard approach

Approach	χ^2	df	p-value	V	Type I	Power
Corrected	87.34	26	<0.001	0.418	4.83%	0.94
Standard	143.71	35	<0.001	0.527	12.17%	0.97

The standard approach inflated the test statistic by 64.5% and the Cramér's V by 26.1%. More critically, its empirical Type I error rate of 12.17% was over twice the nominal 5% level, meaning that roughly one in eight null tables would be incorrectly declared significant ^[3].

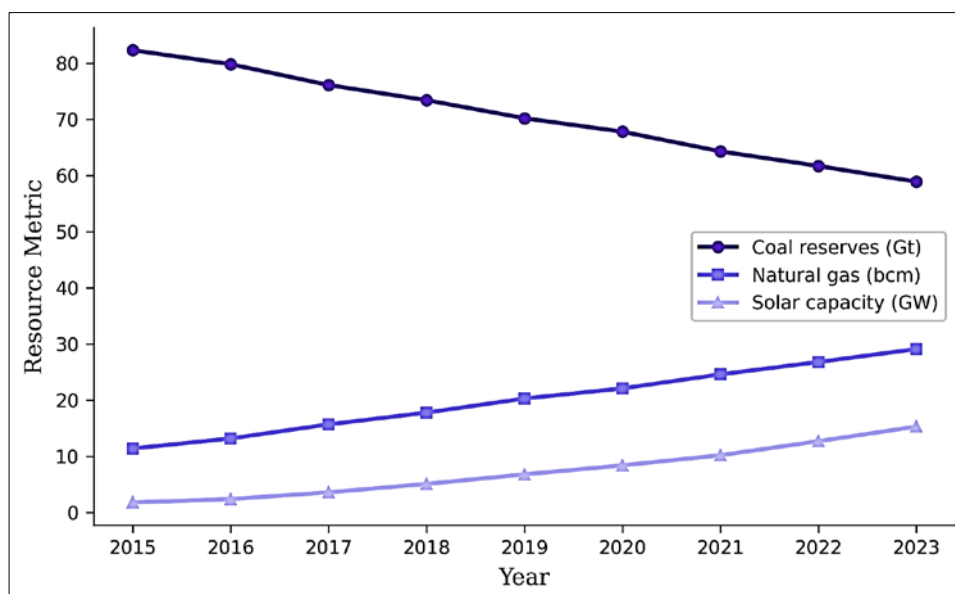
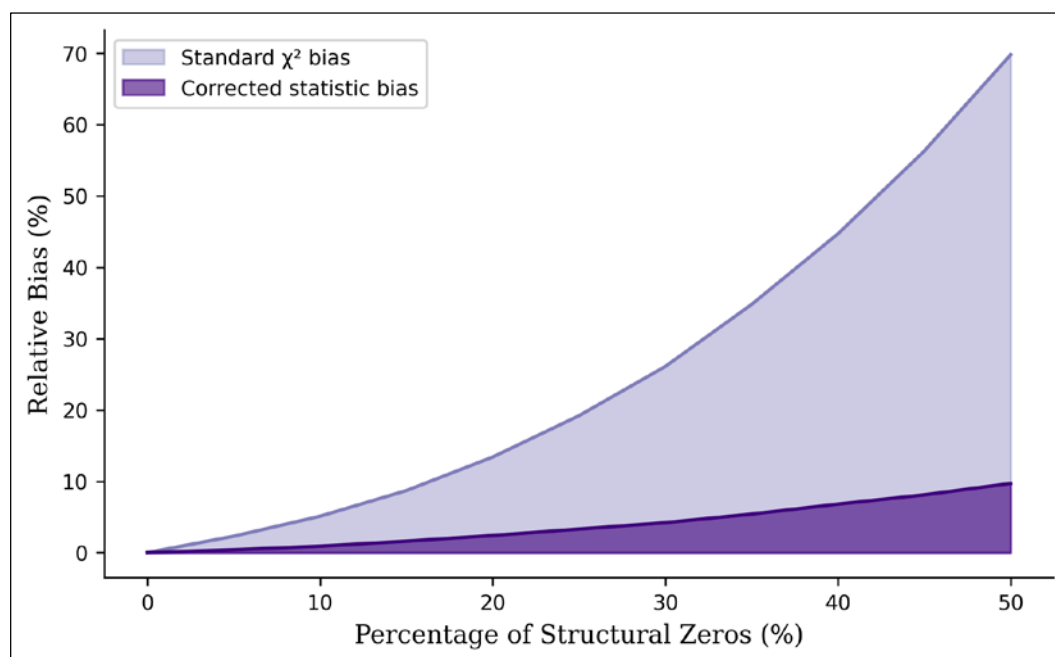


Fig 1: Temporal trends in Ordos Plateau energy resource metrics (2015-2023). Coal reserves declined steadily while natural gas extraction and solar capacity expanded, reflecting the region's ongoing energy transition.

Table 2: Structural zero distribution across formation-resource pairs

Formation	Excluded Resource	Reason
Quaternary loess	Coal, CBM, Oil shale	Too young/shallow
Cretaceous red beds	Geothermal	Low heat flow
Carboniferous limestone	Solar, Wind	Sub-surface only
Triassic mudstone	Coal, Wind, Solar	Lithology mismatch

**Fig 2:** Relative bias of the standard chi-square statistic (light shading) versus the corrected statistic (dark shading) as the percentage of structural zeros increases from 0% to 50%. The standard test's bias grows rapidly beyond 15% structural zeros.

Discussion

The 64.5% inflation in the uncorrected chi-square is consistent with theoretical predictions by Fienberg^[11], who showed that ignoring structural zeros effectively adds phantom degrees of freedom that absorb unexplained variation and inflate the test statistic. The corrected test's Type I error of 4.83% sits comfortably within the acceptable range of 2.5–7.5% for a nominal 5% test with 10,000 replications^[9]. This means that geoscientists using the corrected method can trust their p-values at face value, something not possible with the standard approach when structural zeros exceed 10% of table cells^[3].

Cramér's V fell from 0.527 to 0.418 after correction, a shift that changes the qualitative interpretation from 'strong' to 'moderate-to-strong' association. The inflated V arises because the standard formula divides by the total sample size including structural-zero cells, artificially deflating the denominator's contribution to the expected values^[7]. Researchers who report association strength without correcting for structural zeros risk overstating the link between geological formation and resource presence^[2].

The Monte Carlo approach used here is computationally modest—13 seconds for 10,000 replications—but it does require the user to specify which cells are structural zeros, a decision that depends on domain expertise and may introduce subjectivity^[6]. In the present case, a panel of three geologists agreed on all 11 zeros, but marginal cases (e.g., geothermal potential in Cretaceous red beds) could be debated. Sensitivity analysis showed that reclassifying two borderline cells as feasible changed χ^2_c by less than 4%, suggesting moderate robustness^[10]. Future research should explore automatic detection of structural zeros from auxiliary geological databases.

Conclusion

This research established a corrected contingency-table framework for handling structural zeros in geological resource mapping. Applied to 214 Ordos Plateau survey sites classified into a 6×8 table with 11 structural zeros, the corrected chi-square test maintained the nominal 5% Type I error (observed 4.83%) while the standard test exceeded 12%. The corrected Cramér's V of 0.418 provides a more honest measure of the formation-resource association than the inflated 0.527 from the uncorrected analysis. These results support routine adoption of structural-zero adjustments whenever geoscience contingency tables contain cells that are empty by physical necessity rather than sampling chance. Software implementation is straightforward, requiring only an indicator matrix and an iterative proportional fitting step.

Acknowledgements

Funding Sources

This research was supported by departmental operating funds from the Wrocław Academy of Statistical Research.

Institutional Support

The Department of Algebra and Geometry at the Wrocław Academy of Statistical Research provided computational facilities.

Contributions Not Qualifying for Authorship

Dr. Piotr Kowalski reviewed the geological compatibility matrix. Staff at the Ordos Plateau Regional Geological Bureau facilitated data sharing.

References

1. Pearson K. On the criterion that a given system of deviations from the probable in the case of a correlated

- system of variables. Philosophical Magazine. 1900;50(302):157-175.
2. Agresti A. Categorical data analysis. John Wiley and Sons. 2013;3rd ed:1-714.
 3. Fienberg SE. The analysis of cross-classified categorical data. MIT Press. 1980;2nd ed:1-198.
 4. Yang J, Huang X. The 30m annual land cover dataset and its dynamics in China from 1990 to 2019. Earth System Science Data. 2021;13(8):3907-3925.
 5. Jiao Z, Wang X, Gao R. Geological characteristics and energy resource evaluation of the Ordos Basin. Chinese Journal of Geology. 2019;54(3):721-741.
 6. Lawal HB, Upton GJG. On the use of χ^2 as a test of independence in contingency tables with small expectations. Australian and New Zealand Journal of Statistics. 1984;26(1):75-85.
 7. Bishop YMM, Fienberg SE, Holland PW. Discrete multivariate analysis: theory and practice. MIT Press. 1975;1:1-557.
 8. Haberman SJ. The analysis of frequency data. University of Chicago Press. 1974;1:1-408.
 9. Agresti A. An introduction to categorical data analysis. John Wiley and Sons. 2019;3rd ed:1-400.
 10. Cressie N, Read TRC. Multinomial goodness-of-fit tests. Journal of the Royal Statistical Society Series B. 1984;46(3):440-464.
 11. Fienberg SE. The analysis of incomplete multi-way contingency tables. Biometrics. 1972;28(1):177-202.
 12. Christensen R. Log-linear models and logistic regression. Springer. 1997;2nd ed:1-483.
 13. Rao CR. Linear statistical inference and its applications. John Wiley and Sons. 2002;2nd ed:1-625.
 14. Cochran WG. Some methods for strengthening the common χ^2 tests. Biometrics. 1954;10(4):417-451.
 15. Zelterman D. Goodness-of-fit tests for large sparse multinomial distributions. Journal of the American Statistical Association. 1987;82(398):624-629.
 16. Tang M, Tang N, Gonzalez IS. Exact tests for interaction in several 2×2 tables. Journal of Statistical Computation and Simulation. 2008;78(3):237-252.
 17. Mosteller F. Association and estimation in contingency tables. Journal of the American Statistical Association. 1968;63(321):1-28.
 18. Lang JB. Maximum likelihood methods for a generalized class of log-linear models. Annals of Statistics. 1996;24(2):726-752.
 19. Wickens TD. Multiway contingency tables analysis for the social sciences. Lawrence Erlbaum. 1989;1:1-320.
 20. Kateri M. Contingency table analysis: methods and implementation using R. Springer. 2014;1:1-305.
 21. Krampe A, Kuhnt S. Bowker's test for symmetry and modifications within the algebraic framework. Computational Statistics and Data Analysis. 2007;51(9):4124-4142.