Analysis of Simpson's Paradox and Its Applications

Zhihao Lei*

Department of Mathematics, University of Edinburgh, Edinburgh, The United Kingdom

*Corresponding author: s2152884@ed.ac.uk

Abstract. As a matter of fact, the enigmatic nature and far-reaching consequences of Simpson's Paradox have captivated both scholars and professionals. On this basis, this investigation seeks to offer an exhaustive review of the paradox, examining its historical background, mathematical formulations, and its utilization in diverse sectors. To be specific, the study probes into how the paradox influences the efficacy of healthcare treatments, its contribution to sustaining gender inequality in educational settings like UC Berkeley, and the challenges it presents in analyzing data with multiple dimensions. The research also points out existing gaps in the literature, specifically the absence of methodical solutions for addressing the paradox in intricate, multi-faceted scenarios. Overall, the significance of this study lies in its multi-faceted exploration of Simpson's Paradox. It serves as a foundational text that aims to improve statistical literacy and promote a nuanced understanding of data interpretation, thereby contributing to better decision-making across multiple disciplines.

Keywords: Simpson's paradox; statistical phenomena; multidimensional data analysis.

1. Introduction

For years, Simpson's Paradox has been a source of confusion and intrigue for researchers, statisticians, and data analysts [1-5]. The paradox manifests when a clear trend within individual groups vanishes or reverses upon merging those groups. Initially formalized by British statistician Edward H. Simpson in 1951 [6], the concept had already caught the attention of other statisticians like Karl Pearson, who had started investigating the intricacies of statistical interpretation as far back as the late 19th century [7]. The paradox has been a subject of extensive research and discussion across various academic disciplines, including statistics, epidemiology, and social sciences [8]. It serves as a cautionary tale in the interpretation of statistical data, emphasizing the importance of considering confounding variables and the dangers of drawing conclusions based solely on aggregate data [9]. Over the years, the paradox has been examined from multiple perspectives, including its mathematical underpinnings, its impact on public policy, and its ethical implications [10, 11]. The paradox has also been studied in the context of machine learning and data science, where it poses challenges in the interpretation of complex, multidimensional data [12].

The study of Simpson's Paradox has evolved significantly over the years, with numerous examples emerging in various fields. In medicine, the paradox has been observed in clinical trials, where it can significantly affect the interpretation of treatment effectiveness [13]. Economists have also encountered the paradox when analyzing GDP and per capita GDP growths, leading to potentially misleading conclusions about economic progress [14]. Latest progress in the domain has centered on pinpointing the factors that can result in deceptive epidemiological and statistical findings because of the paradox [10]. Scholars have suggested models and loss functions for assessing the correlation between two variables within various subgroups, offering a more detailed comprehension of the paradox. These advancements reveal that although the paradox is widely acknowledged, it remains a vibrant field of ongoing study, consistently yielding new perspectives and techniques.

The motivation for writing this dissertation stems from the persistent relevance and complexity of Simpson's Paradox in contemporary research. Despite extensive studies and the development of methodologies to mitigate its effects, the paradox continues to confound researchers and mislead interpretations. This dissertation aims to provide a comprehensive overview of the current state of research on Simpson's Paradox, its implications in various fields, and potential solutions to mitigate its effects. On this basis, it is hoped to contribute to a clearer understanding of this intricate phenomenon and guide future research in tackling its challenges. The dissertation will delve into the mathematical descriptions, real-world applications, limitations, and future outlooks concerning Simpson's Paradox. The dissertation aims to serve as a comprehensive guide for both academics and practitioners, providing insights into the mathematical descriptions, real-world applications, limitations, and future outlooks concerning Simpson's Paradox. By offering a multi-faceted exploration of this complex phenomenon, the dissertation aims to improve statistical literacy, promote a nuanced understanding of data interpretation, and contribute to better decision-making across multiple disciplines.

2. Basic Descriptions

Simpson's Paradox refers to a statistical oddity in which a trend noticeable within distinct groups flips when those groups are amalgamated. Put more simply, it's a scenario where the overall data portrays a narrative that diverges from what the data from each individual group suggests [15-18]. Mathematically, Simpson's Paradox often occurs in the context of 2×2 contingency tables. Let's consider two groups A and B with variables x and y. In both groups, x increases with y, but when the data from A and B are combined, x decreases with y [16, 19]. The paradox can be represented using conditional probabilities. Let $P(A|B)$ and $P(A|C)$ be the conditional probabilities of event A given B and C respectively. The paradox occurs when: $P(A|B) > P(A|C)$ and $P(A|B') > P(A|C')$ but, $P(A|B \cup B') < P(A|C \cup C')$. This mathematical description is crucial for understanding the paradox's implications in various fields, from healthcare to policy-making [17, 18]. Understanding the mathematical intricacies of Simpson's Paradox is not just an academic exercise; it has real-world implications. Incorrect interpretations due to the paradox can lead to wrong conclusions and decisions, especially in fields like healthcare and information systems [17, 18]. By dissecting the mathematical framework behind Simpson's Paradox, this dissertation aims to provide a comprehensive understanding that can aid in more accurate data interpretation and decision-making.

3. Situations and Applications

3.1. UC Berkeley Gender Bias

One of the most well-known examples of Simpson's Paradox is the case involving gender bias in graduate admissions at UC Berkeley. At first glance, the university appeared to be discriminating against women, as the general admission rates for men were notably greater than those for women [3, 19]. However, a more detailed analysis of the data presented a different narrative, serving as a textbook case of Simpson's Paradox. The paradox becomes evident when scrutinizing the admissions statistics at the level of individual departments. Contrary to the overall trend, specific departments exhibited either no gender bias or a slight preference for admitting women [3, 20]. The underlying cause of this paradoxical situation was that women tended to apply to more competitive departments with lower acceptance rates, whereas men were more likely to apply to departments that were less competitive and had higher acceptance rates [3]. This case is often cited in discussions about statistical procedures and their implications in social sciences [15, 20]. The inappropriate application of statistical tests, such as the $\chi 2$ test, can lead to misleading conclusions about gender bias [21].

Table 1. Data description.

	All		Men		Women	
	Applicants	Admitted	Applicants	Admitted	Applicants	Admitted
Total	12,787	42%	8,375	46%	4,412	33%

The UC Berkeley case serves as a cautionary tale about the importance of understanding the data and the statistical methods used for analysis. It highlights the need for a nuanced approach to interpreting aggregate data, especially when it comes to sensitive issues like gender bias [19, 21]. Understanding Simpson's Paradox can help in avoiding erroneous conclusions and in formulating more effective policies for gender equality. It also serves as an educational tool for teaching statistical reasoning and for fostering a more nuanced understanding of how bias can manifest in complex ways [22, 23]. The data is given in Table. 1 and Table. 2.

Table 2. Data analysis.

Department	All		Men		Women	
	Applicants	Admitted	Applicants	Admitted	Applicants	Admitted
A	949	66%	837	64%	112	85%
B	590	61%	553	61%	37	66%
C	922	34%	346	36%	576	33%
D	823	36%	439	34%	384	38%
E	585	24%	182	27%	403	22%
F	718	8%	363	7%	355	9%
Total	4587	39%	2720	45%	1867	31%

Notes: higher proportion of applicants from one gender succeeded compared to the other; higher number of applicants compared to the opposite gender; bold the two departments that received the most applications from each gender

3.2. Kidney Stone Treatment

Simpson's Paradox has found a notable application in the field of healthcare, specifically in the treatment of kidney stones. Initially, treatment methods were evaluated based on their overall success rates. However, when Simpson's Paradox was considered, the conclusions about the most effective treatment method changed dramatically [24, 25]. The paradox arises when the success rates of different treatment methods for kidney stones are evaluated. For example, Treatment A might appear more effective when considering the overall population but less effective when the data is broken down into subgroups based on the size of the kidney stones [24, 25]. This discrepancy can occur due to various confounding variables such as the size and type of kidney stones, and the age and health condition of the patients [25].

Table 3. Data for kidney stone treatment.

Treatment Stone size	Treatment A	Treatment B
Small stones	Group 1 92% (83/90)	Group 2 85% (240/283)
Large stones	Group 3 74% (204/276)	Group 4 73% (66/91)
Both	78% (287/366)	82% (306/374)

Note: Legend, the treatment with higher success rate

The paradox became especially clear in a historical analysis of the effectiveness of kidney stone removal procedures. When the size of the kidney stones was considered, the success rates appeared to shift, prompting a reassessment of the role of open surgery as a treatment option [25]. Understanding Simpson's Paradox in the context of kidney stone treatment is crucial for medical practitioners and policy-makers. It ensures that the treatment methods are evaluated more rigorously, taking into account various confounding variables [24, 25]. This nuanced approach can lead to more effective and personalized treatment plans for patients. Additionally, the paradox acts as a measure of reliability for statistical tests employed in assessing treatment approaches, like the Wilcoxon-Mann-Whitney (WMW) rank sum test [24]. As a result, being aware of the paradox can substantially enhance the caliber of healthcare delivery and lead to improved patient results. The results is given in Table 3.

3.3. Multidimensional Data Analysis

Simpson's Paradox is not just confined to simple two-dimensional data; it also manifests in multidimensional spaces. This application is particularly relevant in the era of big data, where datasets often have multiple dimensions. The paradox can significantly impact the interpretation of such data, leading to potentially harmful or misleading conclusions [1, 2]. The paradox becomes evident in the analysis of multidimensional data when exploring the connections between different characteristics. A straightforward yet effective method has been devised to automatically pinpoint every occurrence of Simpson's Paradox, which is influenced by diverse sub-groups and separating attributes in a multidimensional dataset [1]. The paradox can alter various aspects like the dimensionality, the size of the sub-groups, the involvement of individual data points, and more [1].

This intricate nature of the paradox opens the door to a range of fascinating but hitherto unexplored questions concerning the distribution of instances of the multidimensional Simpson's Paradox [1]. For instance, the paradox can have a significant detrimental effect on big data, leading to incorrect conclusions that could result in harmful outcomes [2]. Grasping the nuances of Simpson's Paradox in the realm of multidimensional data is vital for data scientists, statisticians, and researchers. It ensures a more rigorous evaluation of the data, considering various confounding factors [1, 2]. Such a detailed approach can result in more precise interpretations and improved decision-making, particularly in sectors that are highly dependent on multidimensional data like healthcare, finance, and social sciences [1, 2]. Moreover, the paradox serves as a robustness check on statistical tests used in multidimensional data analysis. Therefore, recognizing the paradox can significantly impact the quality of data interpretation and contribute to better outcomes in various applications [1, 2].

4. Limitations and Prospects

Despite its widespread recognition and applications across various fields, the study of Simpson's Paradox is not without its limitations. One of the most significant challenges is the paradox's counterintuitive nature, which often leads to misinterpretation of data, especially among those who are not well-versed in statistical reasoning [1, 2]. Furthermore, the existing literature primarily focuses on identifying and analyzing the causes of the paradox but lacks a systematic approach to resolving it in multidimensional spaces [2]. This gap is particularly concerning given the increasing complexity of data in the era of big data and machine learning. Additionally, while the paradox has been studied in specific contexts like healthcare, gender bias, and research funding, there is a lack of comprehensive research that explores its implications in other equally important areas such as environmental science, economics, and public policy [3, 4].

The future of research on Simpson's Paradox holds promising avenues for exploration. One of the most exciting prospects is the development of machine learning algorithms specifically designed to detect and possibly resolve instances of the paradox in complex, multidimensional datasets [1, 2]. Such advancements could revolutionize the way data is interpreted across various disciplines. Moreover, there is a growing interest in understanding the ethical implications of the paradox, especially in decision-making processes that affect human lives, such as medical treatments and social policies [5]. This ethical dimension opens doors for interdisciplinary research involving statisticians, ethicists, and policymakers. Finally, as data continues to grow in complexity, there is an urgent need for educational initiatives that aim to improve statistical literacy, focusing on the understanding and interpretation of paradoxical phenomena like Simpson's Paradox [4, 5].

5. Conclusion

To sum up, Simpson's Paradox is a statistical phenomenon that has far-reaching implications across various fields, from healthcare and gender bias to multidimensional data analysis. This study has delved into the origins, mathematical descriptions, and diverse applications of the paradox. It has highlighted how the paradox can lead to misleading conclusions in healthcare treatment effectiveness, perpetuate gender bias in educational institutions, and complicate the interpretation of multidimensional data. However, the study acknowledges the limitations in the current body of research, particularly the lack of systematic approaches to resolving the paradox in complex, multidimensional spaces. Future research promises advancements in machine learning algorithms designed to detect and resolve instances of the paradox, as well as interdisciplinary studies focusing on its ethical implications. The significance of this research lies in its comprehensive exploration of Simpson's Paradox, serving as a foundational text for both academics and practitioners. It aims to improve statistical literacy and promote a nuanced understanding of data interpretation, thereby contributing to better decision-making across multiple disciplines.

References

[1] Sharma R, Garayev H, Kaushik M, et al. Detecting Simpson’s Paradox: A Machine Learning Perspective. International Conference on Database and Expert Systems Applications. Cham: Springer International Publishing, 2022: 323-335.

[2] Xu J, Pei J, Cong Z. Finding Multidimensional Simpson's Paradox. ACM SIGKDD Explorations Newsletter, 2022, 24(2): 48-60.

[3] Albers C J. Dutch research funding, gender bias, and Simpson’s paradox. Proceedings of the National Academy of Sciences, 2015, 112(50): E6828-E6829.

[4] Swamy P, Mehta J S, Tavlas G S, et al. Two applications of the random coefficient procedure: Correcting for misspecifications in a small area level model and resolving Simpson's paradox. Economic Modelling, 2015, 45: 93-98.

[5] Bandyopadhyay P S, Raghavan R V, Dcruz D W, et al. Truths about Simpson’s Paradox: Saving the paradox from falsity. Logic and Its Applications: 6th Indian Conference, ICLA 2015, Mumbai, India, January 8-10, 2015. Proceedings 6. Springer Berlin Heidelberg, 2015: 58-73.

[6] Woo E J. Pearls: Simpson’s Paradox—Understanding Numbers That Don’t Seem to Make Sense. Clinical Orthopaedics and Related Research, 2019, 477(11): 2427.

[7] Spanos A. Yule–Simpson’s paradox: the probabilistic versus the empirical conundrum. Statistical Methods & Applications, 2021, 30: 605-635.

[8] Goltz H H, Smith M L. Yule-Simpson's paradox in research. Practical Assessment, Research, and Evaluation, 2010, 15(1): 15.

[9] Rojanaworarit C. Misleading epidemiological and statistical evidence in the presence of Simpson's paradox: an illustrative study using simulated scenarios of observational study designs. Journal of Medicine and Life, 2020, 13(1): 37.

[10] Wang J, He J, Xu W, et al. Learning to Discover Various Simpson's Paradoxes. Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. 2023: 5092-5103.

[11] Rücker G, Schumacher M. Simpson's paradox visualized: the example of the rosiglitazone meta-analysis. BMC medical research methodology, 2008, 8(1): 1-8.

[12] Cani P D, Depommier C. Reply to ‘Simpson’s paradox in proof-of-concept studies’. Nature medicine, 2019, 25(11): 1640-1641.

[13] Ma Y Z. Simpson’s paradox in GDP and per capita GDP growths. Empirical Economics, 2015, 49(4): 1301-1315.

[14] Zhang B, Heng S, MacKay E J, et al. Bridging preference-based instrumental variable studies and cluster-randomized encouragement experiments: Study design, noncompliance, and average cluster effect ratio. Biometrics, 2022, 78(4): 1639-1650.

[15] Streiner D L. Statistics Commentary Series. Commentary No. 38: Simpson's Paradox and the Mantel-Haenszel χ2. Journal of Clinical Psychopharmacology, 2020, 40(2): 109-111.

[16] Sarkar P, Bandyopadhyay P S. Simpson's Paradox: A Singularity of Statistical and Inductive Inference. arXiv preprint arXiv:2103.16860, 2021.

[17] Chipman J, Braun D. Simpson's paradox in the integrated discrimination improvement. Statistics in medicine, 2017, 36(28): 4468-4481.

[18] Kock N, Gaskins L. Simpson's paradox, moderation and the emergence of quadratic relationships in path models: an information systems illustration. International Journal of Applied Nonlinear Science, 2016, 2(3): 200-234.

[19] Mittal Y. Homogeneity of subpopulations and Simpson's paradox. Journal of the American Statistical Association, 1991, 86(413): 167-172.

[20] Chu K H, Brown N J, Pelecanos A, et al. Simpson's paradox: A statistician's case study. Emergency Medicine Australasia, 2018, 30(3): 431-433.

[21] Selvitella A. The ubiquity of the Simpson's Paradox. Journal of Statistical Distributions and Applications, 2017, 4(1): 1-16.

[22] Borenstein M, Hedges L V, T. H J P, et al. Chapter 33: Simpson's Paradox. Introduction to meta-analysis. Hoboken, NJ: John Wiley & Sons, Inc., 2021.

[23] Nee J, Macfarlane Smith G, Sheares A, et al. Advancing social justice through linguistic justice: Strategies for building equity fluent NLP technology. Equity and Access in Algorithms, Mechanisms, and Optimization. 2021: 1-9.

[24] Lu C. Causal Confirmation Measures: From Simpson's Paradox to COVID-19. Entropy, 2023, 25(1): 143.

[25] Julious S A, Mullee M A. Confounding and Simpson's paradox. Bmj, 1994, 309(6967): 1480-1481.