Expanded analysis of Simpson's Paradox: A Case Study

Xinyan Wang

No.11 high school, Wuhan, China

bettywang@ldy.edu.rs

Abstract. Simpson' s paradox is a famous paradox in economic field. It explains the abnormal statistical results with respect to two different variables. That is to say, in some kinds of census, the trend of the whole group may not be able to represent the separate groups. The application of it includes enrollment of universities, success rate of treatment, the effect of smoking on life span, admissions policies and so on. The paradox plays a significant role in statistical analysis, because in some terms, it can decide the accuracy of the data sets. The paper aims to seek possible solutions to reduce the uncertainties emerged in the paradox with respect to mathematical methods. In this paper, theoretical and real case analysis are presented. Several representative cases are analyzed, and previous methods are included. The author raises an improvement for the methods above. The conclusion is that if proper weight is added, there will be little error in spite of the occurrence of Simpson' paradox.

Keywords: Simpson' paradox, causality, weighted average.

1. Introduction

When it comes to paradox, very limited number of people know that this word is derived from a Greek word "*paradoxon*", which refers to something that is not in line with common sense. Simpson' s paradox is a famous paradox in the field of statistical science discovered by E.H. Simpson in 1951. Scholars have already done researches on this issue because it has a wide application in many fields, such as the tax policy and the acceptance rate of universities. Some people have already used Simpson' s paradox in areas like the ranking of universities [1]. And the finding is that it can hardly affect the ranking. Because the evaluation indicators are more than one. Only technical analysis indicators can have Simpson' s paradox, but the quantity is relatively small. Therefore, the effect is very limited. However, in more accurate analysis, it can affect the accuracy of statistical analysis, so people are trying to minimum the uncertainty. This paper aims at reviewing the researches done by former scholars and propose a more convenient method to reduce the uncertainty. This method is suggested by having a relatively deep understanding of the essence of Simpson' s paradox and summarizing previous methods. It is found that the cause of Simpson' s paradox is the irrelevant variables. This means if we choose the proper variables and put them into different groups, the result may be more accurate. In the second section, the paper will summarize the discoveries and attempts made by scholars as well as the numerous applications of the renowned paradox. That means a macroscopic view is offered. In the third section, the background of the case studied along with the improved method are provided. In the fourth section, statistics reveal the essence of the paradox. Based on this, giving the samples different weight can make the result of each group consistent with the total group. The study makes some improvement of the previous methods to raise a more simple and practical solution to minimum the uncertainty that may occur. This will be proved by a well-known real case. More importantly, there is a whole new promising field where the paradox can be applied to suggested in the paper. Further study is appreciated.

2. Literature Review

2.1. Mathematical Expressions of Simpson' s Paradox

E.H. Simpson first propose Simpson's paradox in 1951[2]. In this paper, he referred to a test which aimed to examine whether in a pack of cards the proportion of court cards (King, Queen, Knave) was

associated with color. Some cards played by babies were dirty, so the investigator added "dirty" to the classification in case it is relevant. He also deduced a positive association between redness and plainness both among the dirty cards and among the clean, but the result did not suggest that association. From this, Simpson' s paradox is found. The finding suggested that irrelevant variables may lead to confusion.

Some scholars such as Geng Zhi found that the cause of Simpson' s paradox is causality [3]. This means some irrelevant variables may affect the accuracy of a test. Let us take a look at table 2. Different departments have different preference for gender. If we assume each department has the same standard for enrollment, the statistic of the whole school (table 1) can represent the outcome of every department. However, each department has different standard for enrollment, a factor that cannot be ignored. The statistic of each department cannot represent the final outcome of the whole school. Therefore, it is crucial to decide which factors should be considered.

Former scientists have proposed several ways to reveal Simpson' s paradox systematically. For instance, Alipourfard et al. [4] assumed that there are a set of input $X = \{X_1, X_2, \dots, X_m\}$ and an output Y. Then the solution has three steps. First, according to some covariateX_c, it separates data into many homogenous subgroups. Second, a linear analysis is used to observe trends with other covariateX_j respectively, inside the subgroups and the aggregate data. At last, it evaluates the performance of the models to describe the dis-aggregated data in comparison with aggregate data to recognize the crucial disaggregation."

First, a linear regression is used.

$$E[Y|X_j = x_j] = f(\alpha + \beta X_j) \tag{1}$$

Then, it is combined with conditional expectation.

$$E[Y|X_j = x_j, X_c = x_c] = f(\alpha(xc) + \beta(xc)X_j) \tag{2}$$

Later, Xiayi Wang define a characteristic Simpson pair (X_j, X_c)[5].

By comparing the sign of β, she is able to find Simpson pair, which is able to eliminate or reverse the trend of $X_j$and Y. From this we can see that we no longer need to recognize Simpson' s paradox by calculate three sets of data. We can simply use the linear regression formula above and observe the sign of β.

WANG B, WU P, KWAN B, et al. studied Simpson's paradox related with conditional expectation, categorical data analysis and time-to-event data analysis [6-8]. They tried to address the problem in continuous data.

If:

$$a/b = c/d, \text{ then } a/b = (a + c)/(b + d) = c/d. \tag{3}$$

But is it also a proper explanation for inequalities of fractions? Specifically, assume $s_{ij}, n_{ij} (i = 1,2, j = 1,2)$ are positive numbers with $s_{1j}/n_{1j} < s_{2j}/n_{2j}, j = 1,2$. Is it true that $(s_{11} + s_{12})/(n_{11} + n_{12}) < (s_{21} + s_{22})/(n_{21} + n_{22})$? Let us look at this case: 3/4 < 7/9 and 2/3 < 15/22, but (3+2)/ (4+3) =5/7 > 22/31 = (7+15)/ (9+22). The pooled data shows a relation opposite to what they expected. Let Y be a random variable with $E|Y| < \infty$. Suppose $X_1$and $X_2$are two random variables with$X_i \in \{1,2, \dots k_i\}$, where $k_i (\geq 2)$, i=1,2 are positive integers. Then, for any $m \in \{1, \dots, k_1\}$, then

$$E[Y|X_1 = m] = \sum_{n=1}^{k_2} E[Y|X_1 = m, X_2 = n] \, Pr\{X_2 = n|X_1 = m\} \tag{4}$$

Table 1: Average scores of male and female students in two schools

School (X_1)	Male (1)		Female (2)	
	n	Average	n	Average
Alpha (1)	80	84	20	80
Beta (2)	20	85	80	81

$E[Y|X_1 = 1, X_2 = 1] = 84,\ E[Y|X_1 = 1, X_2 = 2] = 80,$
$E[Y|X_1 = 2, X_2 = 1] = 85,\ E[Y|X_1 = 2, X_2 = 2] = 81,$
$Pr\{X_2 = 1|X_1 = 1\} = 0.8,\ Pr\{X_2 = 2|X_1 = 1\} = 0.2,$
$Pr\{X_2 = 1|X_1 = 2\} = 0.2,\ Pr\{X_2 = 2|X_1 = 2\} = 0.8$
It is obvious that

$$E[Y|X_1 = 1, X_2 = n] < E[Y|X_1 = 2, X_2 = n], n = 1,2. \quad (5)$$

Equation (5) illustrates that students in both genders in School Beta receive higher scores. While they are calculating the average score with respect to two schools, it is necessary to take the gender proportion in to account. And in (4) the average grades of schools are obtained from weighted average of the grades of males and females. In all, the imbalance of the gender distribution in two schools is the main reason for this phenomenon to occur.

The research above has some similarities, they all try to find the decisive factor for Simpson's paradox to occur. Alipourfard used linear regression to explain the essence of Simpson' s paradox. Xiayi Wang defined Simpson pair to recognize Simpson' paradox. Geng Zhi suggested a way based on causality to avoid the inaccuracy of Simpson' s paradox. WANG B, WU P, KWAN B, et al. showed how this phenomenon can occur for continuous, categorical and survival outcome. Some suggested that Simpson' s paradox can be avoided by downgrading, correction, and deletion. These works reveal the mathematical essence of Simpson' s paradox or give general solutions, yet they are not trying to solve it directly.

2.2. Wide Applications of Simpson' s Paradox

There are studies suggesting that the paradox can bring benefits to areas regarding medical research, artificial intelligence, etc. Here are the examples. A famous study demonstrated that it could be used to compare the treatments of kidney stones [9,10]. To be more specific, when the diameters of the stones were not considered, percutaneous nephrolithotomy was more favorable, while open surgery was suitable for stones of known sizes. The astonishing discovery depends on the sizes of stones, a confounding variable. Hence, patients can choose to either have open surgery when the stone is large or percutaneous nephrolithotomy when the stone is pretty small. The decision is made more efficient thanks to this study. Another study with respect to computer science [11] introduced a method called classification tree, a popular algorithm in machine learning. It is often used to predict an outcome. It makes it easier for people who make decisions related to the paradox. There is also one study stated that the disadvantages outweigh the advantages in Simpson' s paradox [12]. More specifically, an individual may have many dimensions, but this does not mean that they will change over time. Also, two variables may influence groups of people but not the inner heart of everyone. That is to say, the impact of such method has not reached individual level. Yet, the affect can be partly reduced by conditioning on subgroups. Briefly speaking, there are strengths and drawbacks when using the paradox, however, proper solutions are on their way.

2.3. Inspiration of This Paper

The previous part analyzes the definition, appearance and applications of Simpson' s paradox. Former scholars have already made attempts to explain phenomena and reduce uncertainties using this paradox yet are complicated and don' t cover other intriguing fields. This paper, based on the identification of Simpson' s paradox and further classify statistics into more groups to make the result more accurate, suggests a more convenient and direct method that is related to weighted average and different from the researches above. It is possible to improve accuracy in most of the circumstances. Moreover, this study mentions that the paradox can be used to monitor the psychological health of human beings, which remains a relatively new area.

3. Methodology

Simpson' s paradox refers to two sets of data under a certain condition, each of which satisfies a certain property when discussed separately, but once combined, it may lead to opposite conclusions. The mathematical formula is

$$E(Y|T=1,X=x)=E\,(T=1|X=x)/E(X=x)E(Y,T=1,X=x) \tag{6}$$

The essence of it is the proportion each factor accounts for, so we use methods of separating the variables into different groups and applying weighted average.

$$\bar{x}=\frac{x_1f_1+x_2f_2+\cdots+x_nf_n}{n} \tag{7}$$

Consider the case below.

Table 2. General enrollment data

	All		Men		Women	
	Applicants	Admitted	Applicants	Admitted	Applicants	Admitted
Total	12763	41%	8442	44%	4321	35%

The data sets (table 2 and table 3) are gained from a website [7]. It illustrates the statistics of an experiment carried out by University of California, Berkeley in 1973 with the aim of investigating the acceptance rate of the postgraduates who are classified by their gender. From table 2, it' s obvious that males are more likely to be accepted compared to females.

Table 3. Detailed enrollment data

Department	All		Men		Women	
	applicants	admitted	applicants	Admitted	Applicants	Admitted
A	933	64%	825	62%	108	82%
B	585	63%	560	63%	25	68%
C	918	35%	325	37%	593	34%
D	792	34%	417	33%	375	35%
E	584	25%	191	28%	393	24%
F	714	6%	373	6%	341	7%
Total	4526	39%	2691	45%	1835	30%

However, when looking into some concrete data in table3, it is strange to find that among the 85 colleges, 6 of them have discrimination to males while 4 of them have discrimination to females. Therefore, it is reasonable to find an explanation to this phenomenon.

In fact, when people consider this situation, they always assume that the proportion of male in every department is approximately the same. If so, it is easy to conclude that the university has a preference for male applicants. But this is quite rare in real cases. So different percentages of a certain gender leads to a result that is contrary to the common sense.

The paper tries to separate the variables into different groups, for example, gender and department, like table3, so the relationship between acceptance and gender in different departments can be clearly displayed. When the study gets all the detailed data, all kinds of relationships people are interested in can be studied. Moreover, people can add weight to different departments to gain a general preference for gender in the university. The paper applies the formula

$$\bar{x}=\frac{x_1f_1+x_2f_2+\cdots+x_nf_n}{n} \tag{8}$$

So, the preference of male in the university is 62%×825/2691+63%×560/2691+37%×325/2691+33%×417/2691+28%×191/2691+6%×373/2691= 45%

4. Results

The occurrence of Simpson' paradox is because of different weight of the data. Therefore, it is reasonable to control the variable. The study finds that more groups of data can be convenient for the study of association. The result of calculation using weighted average is in line with that gained from addition and division. From this we learn that separating variables into more groups can help us analyze the relationship between two specific variables we want to study and adding certain weights to various groups can be conducive to the eradication of ambiguity. In general, this method is straightforward and accurate.

5. Conclusion

This paper explains the meaning of Simpson' s paradox, summarize the former research in field and come up with a small improvement, which is separating the variables into more groups and control the variables. The works of previous scholar are abstract and professional. The study' s solution is more direct and easier to carry out. However, this solution has many drawbacks. For example, the groups of variables may be too much, so it is hard to eliminate irrelevant variables. In this case, latter scholars can use computer programing to draw tables of many variables and calculate the rate they want to study. The author hopes this method can be used to illustrate factors related to personal growth, such as academic performance, physical health, mental well-being and so on. As a result, people can fully understand the most decisive factor of the building of characteristic. Hopefully, data enrich people' s understanding of the world.

References

[1] Can the Simpson paradox fundamentally overturn the scientific validity of university rankings——Discussion on Possible Defects and Improvement Strategies in Comprehensive Evaluation Methods [J]. Research on Education Evaluation in Shanghai, 2023, 12(03): 52-57. DOI: 10.13794/j.cnki.shjee.2023.0043.

[2] Eh S. The interpretation of interaction in contingency tables [J]. [2024-08-09].

[3] Geng Zhi. Causal Inference and Simpson Paradox [J]. Forum on Statistics and Information, 2000, (03): 9-12.

[4] Alipourfard N, Fennell P G, Lerman K. Using Simpson's Paradox to Discover Interesting Patterns in Behavioral Data.2018[2024-08-07]. DOI:10.48550/arXiv.1805.03094.

[5] Wang Xiayi. Avoiding Simpson's Paradox in Heterogeneity Processing Effect Analysis [D]. East China Normal University, 2022.DOI: 10.27149/d.cnki.ghdsu.2022.001312.

[6] WANG B, WU P, KWAN B, et al. An Example of Simpson's Paradox [J]. Shanghai Psychiatry, 2018, 30(02): 139-143.

[7] https://blog.csdn.net/bill2766/article/details/122285064[3]Wang Jian, Ji Ran. The Simpson Paradox in Life [J]. Chinese Statistics2023, (08):66-68.

[8] Bonovas, S.; Piovani, D. Simpson's Paradox in Clinical Research: A Cautionary Tale. J. Clin. Med. 2023, 12, 1633. https://doi.org/10.3390/jcm12041633

[9] Charig, C.R.; Webb, D.R.; Payne, S.R.; Wickham, J.E. Comparison of Treatment of Renal Calculi by Open Surgery, Percutaneous Nephrolithotomy, and Extracorporeal Shockwave Lithotripsy. Br.Med.J.1986, 292, 879–882.

[10] Julious, S.A.; Mullee, M.A.Confounding and Simpson's Paradox. BMJ 1994, 309, 1480–1481.

[11] Shmueli, G., Yahav, I. The forest or the trees?Tackling Simpson's Paradox with classifica- tion trees [J]. Production and Operations Management, 2018, 27(4): 696-716.

[12] Kievit, R., Frankenhuis, W.E., Waldorp, L., Borsboom, D. Simpson's paradox in psycholog- ical science:a practical guide [J]. Frontiers in Psychology, 2013, 4: 513-514.