Parabola Volume 50, Issue 1 (2014)

An Analysis of the Mathematical Structure of Simpson's Paradox with Applications

Congwen Xu[1]

Introduction

In 1899, more than a century ago, Karl Pearson, Alice Lee, and Leslie Bramley-Moore [1] discovered an interesting phenomenon, in which it was found that a statistical association between two different groups was reversed when the two groups were combined. This phenomenon was described again by British statistician Udney Yule [2]. In 1951, it was mentioned in a paper by Edward Hugh Simpson [3], and it is now most commonly known as Simpson's Paradox. In some papers, Simpson's Paradox may also be referred to as the Reversal Paradox, the Amalgamation Process, the Yule-Simpson Effect, or the Ecological Paradox [4].

Over the course of history, Simpson's Paradox has led to several significant data misinterpretations, and it currently plays an important role in applied statistics and epidemiology. For example, a study conducted in 1986 by Charig *et al* [5] indicated that percutaneous nephrolithotomy, a surgical procedure used to remove kidney stones, is more successful in kidney stone treatment than comparative nonsurgical procedures. However, it was later observed upon further analysis that Simpson's Paradox was present in the data. Based on this fact, it was determined that nonsurgical procedures were in fact more effective than percutaneous nephrolithotomy. Similarly, in a famous 1988 diabetic cohort study by Gatling *et al* [6], it was concluded that patients with non-insulin dependent diabetes had a higher mortality rate than those with insulin-dependent diabetes. This was shown to be an inaccurate conclusion when Simpson's Paradox was discovered in the data, and it was noted that non-insulin dependent diabetes usually develops only after age 40. Outside of medicine and health, Simpson's Paradox also occurs in practical consumer application. A US report incorrectly suggested that Delta Airlines had better on-time performance than Alaskan Airlines based on misinterpreted data [7].

The paradox can be observed in the following example with hypothetical data on the success rates for a certain treatment used for a given disease:

In the course of this paper, we will constantly refer back to this contingency table, a chart commonly used in statistics to display frequency distribution of variables in a matrix format. Here, gender (male/female) is known as the stratifying variable, a variable that divides data into smaller subgroups based on their relation to the variable itself. Treatment type (A/B) acts as the independent variable, which is the variable that is being manipulated in the study and often considered to be the cause of changes in

[1]Missions San José High School, Fremont, California, USA

Table 1: Contingency table illustrating Simpson's Paradox

	Male		Female		Combined	
Treatment	Dead	Alive	Dead	Alive	Dead	Alive
A	10	90 (90%)	200	200 (50%)	210	290(58%)
B	39	221 (85%)	77	63 (45%)	116	284(71%)

the other variables (i.e., treatment affects survival rate). Survival rate (alive/dead) acts as the dependent variable, which is the variable that exhibits changes in response to the independent variable. We will assume that there do not exist other external variables that may affect the data.

It can be seen from Table 1 that the survival rate of males undergoing treatment A at 90% is greater than that of males in treatment B at 85%. Likewise, the survival rate of females undergoing treatment A at 50% is greater than that of females undergoing treatment B at 45%. The survival rate of patients undergoing treatment A is greater than that of patients undergoing treatment B in each of the subgroups. On the other hand, however, when the males and females are pooled, the survival rate of people overall undergoing treatment A at 58% is less than that of people overall undergoing treatment B at 71%. Simpson's Paradox has occurred, and along with it, two questions arise: How is this possible? Which treatment should be administered to someone of unknown gender?

Traditionally, the answer to these questions lies in an algebraic analysis of Simpson's Paradox. However, in this article, we introduce a new graphical approach to analyzing Simpson's Paradox that facilitates understanding and analysis of the paradox. We base our explanation primarily on a graphical display for contingency tables known as BK-plots, introduced in 2001 by Stuart G. Baker and Barnett S. Kramer[8]. Furthermore, we detail a "resolution" to the paradox by explaining how data can be examined when the paradox is present.

A Closer Look at Simpson's Paradox

We begin by creating the following general 2 2 2 contingency table, where A is the independent variable, B the dependent variable, and C the stratifying variable. Furthermore, $\neg A$ denotes the negation of A.

Table 2: General $2 \times 2 \times 2$ contingency table.

	C		$\neg C$		Combined	
	B	$\neg B$	B	$\neg B$	B	$\neg B$
A	a	b	c	d	$a+c$	$b+d$
$\neg A$	e	f	g	h	$e+g$	$f+h$

Mathematically, for whole numbers, Simpson's Paradox is defined by the following set of inequalities:

$$\begin{gathered}\frac{a}{a+b} > \frac{e}{e+f}, \\ \frac{c}{c+d} > \frac{g}{g+h}, \text{ and} \\ \frac{a+c}{a+b+c+d} < \frac{e+g}{e+f+g+h}.\end{gathered}$$

When the sample size is large, it is often also useful to represent Simpson's Paradox in a probabilistic manner. Namely, Simpson's Paradox may be described in probabilistic form with the following set inequalities:

$$\begin{gathered}P(B|A \wedge C) > P(B|\neg A \wedge C) \\ P(B|A \wedge \neg C) > P(B|\neg A \wedge \neg C), \text{ and} \\ P(B|A) > P(B|\neg A)\end{gathered}$$

Both of the above sets of inequalities are equally accurate and informative and the decision to use either definition of Simpson's Paradox should be made on the basis of the real-world application. Aside from these definitional inequalities, there also exist several methods of graphically and mathematically describing Simpson's Paradox. These are capable of providing valuable insight on the nature of this paradox. For example, Shapiro [9] and Baker [8] both introduce graphical representations of Simpson's Paradox. We choose to use the graphical representation in Baker and Kramer [8] in this paper, because of its simplicity and clean method of display. The amount of effort put into defining and describing Simpson's Paradox reveals the vast interest that has been placed on this paradox. Its incorrect or suboptimal decision consequences must be avoided.

Importance

Understanding Simpson's Paradox and knowing how to make an optimal decision when it occurs is the key to any well-conducted research design. The paradox may occur in diverse research studies in both the natural and social sciences and can greatly affect how data will be analyzed and interpreted. In many cases, the final decision in scenarios involving Simpson's Paradox may involve serious consequences, such as life or death. We are led to ask ourselves, if such data were to be collected in real life, what would we do? The significance of our research becomes evident.

It was shown in an article by Pavlides and Perlman [10] that Simpson's Paradox occurs in approximately 1.67% of contingency tables in which the cell probabilities are randomly and uniformly distributed. Equivalently, this means that about 1 in every 60 randomly generated tables of data will exhibit Simpson's Paradox. The significance of Simpson's Paradox cannot be ignored. Since the paradox occurs in a diversity of fields, including, among others, epidemiology, medicine, behavioural science, social sciences,

and even sports, research on Simpson's Paradox has broad applications. It should be a fundamental concern in the research design phase of all qualitative and experimental studies.

Explaining Simpson's Paradox

BK-plots, first introduced by Stuart G. Baker and Barnett S. Kramer [8], are a graphical approach to visualizing data in contingency tables. The plot is best understood by considering an example. Consider Figure 1, which shows the BK-Plot of the data in Table 1.

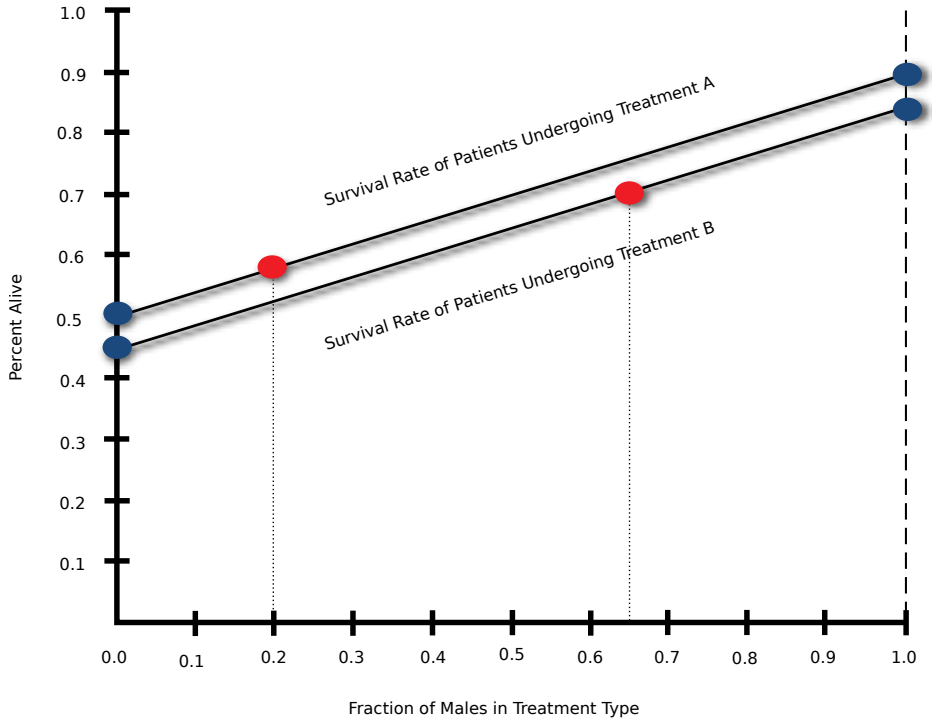

Figure 1: BK-Plot of hypothetical data

As we observe, there are two line segments present in the plot. The upper line segment represents the survival rate of patients undergoing treatment A across different percentages of males and females in the treatment group. For example, the black dot at the very left end of the upper line segment represents the survival rate of treatment A when 0% of the patients are male, or in other words, when all the patients are female. Likewise the grey dot present on the upper line represents the survival rate of treatment A when 20% of the patients were male while the other 80% were female. This resulted in an overall survival rate of 58%, which is the survival rate of the patients in the combined pool. Notice that if there had been more males in treatment A, the grey dot would have shifted higher up on the line segment. In other words, if more of the

patients receiving treatment A had been male, the overall survival rate of the patients in the combined pool would have increased.

While the upper line segment represents survival rates under treatment A, the lower line segment similarly represents the survival rate of patients under treatment B for different percentages of males and females in the treatment group. We see that the two grey points, one on each line, are the survival rates of the combined sample for treatments A and B, while the four black points on the sides represent the individual survival rates of each subgroup: only males undergoing treatment A, only males undergoing treatment B, only females undergoing treatment A, and only females undergoing treatment B. Notice that the treatment A line segment is above the treatment B line segment, indicating that treatment A has a higher survival rate than treatment B in both subgroups, males and females. Furthermore, notice that the grey point on the treatment A line segment is actually lower than the grey point on the treatment B line segment, indicating that the fraction of people alive in the combined sample undergoing treatment A is actually less than that in the combined sample of treatment B. From this, we can conclude that in a BK-Plot, Simpson's Paradox will only occur when one of the line segments is above the other. Furthermore, the grey dot on the higher line segment must be located below the grey dot on the lower line segment even though the upper line is completely above the lower line.

With this knowledge of BK-Plots, we now proceed to use it to analyze Simpson's Paradox. To facilitate this process, we introduce a more generic BK-plot, also showing an instance of Simpson's Paradox (Figure 2).

By looking at the two BK-Plots, several observations can be made. First, notice that in the BK-plots, the grey dots are never directly above or below one another. If this was the case, the grey dot on the higher line would be above the grey dot on the lower line, and Simpson's Paradox could not have occurred (Figure 3). Thus, the two grey dots are never directly above or below each other, in any instance of Simpson's Paradox. In the context of the example, in order for the grey dots to not be directly above or below each other, there must be a different percentage of males in each treatment type. Thus, in order for Simpson's Paradox to occur, there must be an association between the gender (percentage of males) and the treatment type (A/B). More generally, we may conclude that in order for Simpson's Paradox to occur in any data set, there exists an association between the stratifying variable and the independent variable.

Second, one can observe that the two line segments in the BK-plots were never parallel to the x-axis. If either of the two line segments were parallel to the x-axis, the grey dot on the upper line segment would always be above the grey dot on the lower line segment, preventing the paradox from occurring. This situation is shown in Figure 4.

It can be deduced that in order for a line segment to not be parallel to the x-axis, the black dots at the ends of a line segment must not have the same y value. In terms of the contingency table for each subgroup of the independent variable (for example: treatment A, treatment B), which is represented by a line segment, the corresponding dependent variable must take on a different value in each of the two subgroups of the stratifying variable. In terms of causality, the stratifying variable is associated with

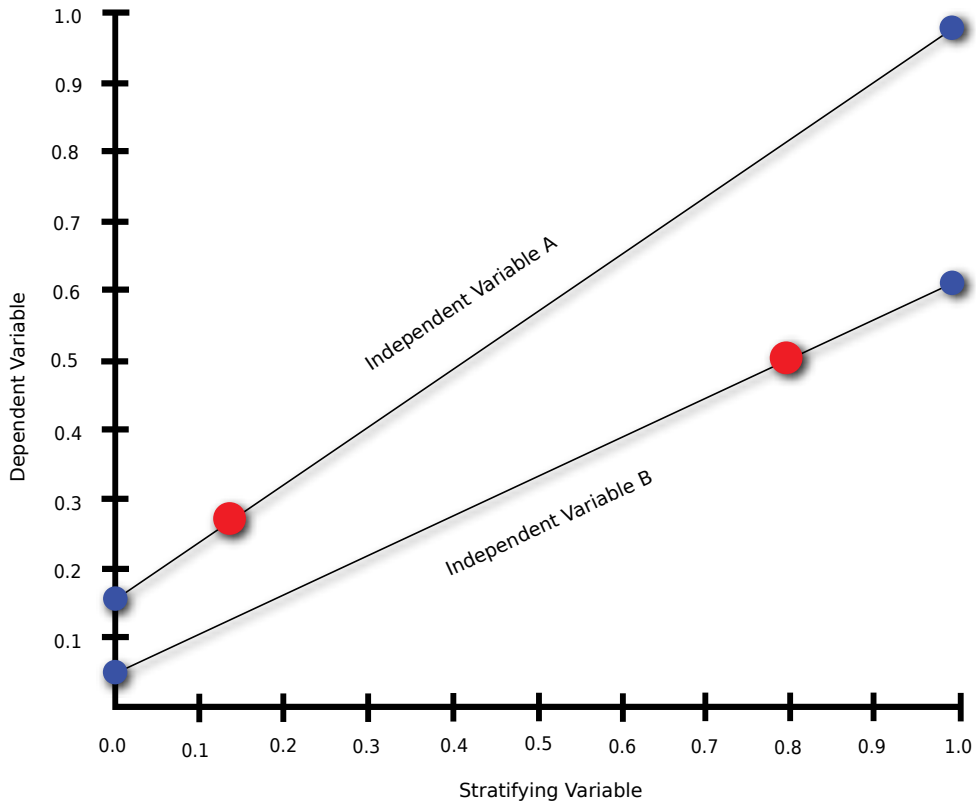

Figure 2: Generic BK-Plot exhibiting Simpson's Paradox

the dependent variable. Given that the stratifying variable must be associated with both the independent variable and the dependent variable when Simpson's Paradox occurs, we conclude that the stratifying variable must then be a confounding variable, an extraneous variable correlated with both the dependent and independent variables. This relationship is visualized in Figure 5.

An arrow pointing from one variable to another variable indicates a causal relationship (i.e., changes in the independent variable affect the dependent variable). It now becomes evident that the underlying cause of Simpson's Paradox is the presence of confounding in the data. We have now successfully explained why the paradox occurs!

Resolving the Paradox

After understanding why Simpson's Paradox occurs, the next step naturally is to try and apply our understanding to data sets. We adopt a method similar to propensity score matching, a method used to reduce bias due to confounding for observational data. To explain how this works, we will use our previous example on treatment types. Without loss of information regarding the association between the independent and dependent variable, we represent all numerical values as cell probabilities (Table 3).

We will consider only the subgroups of the confounding variable (male/female)

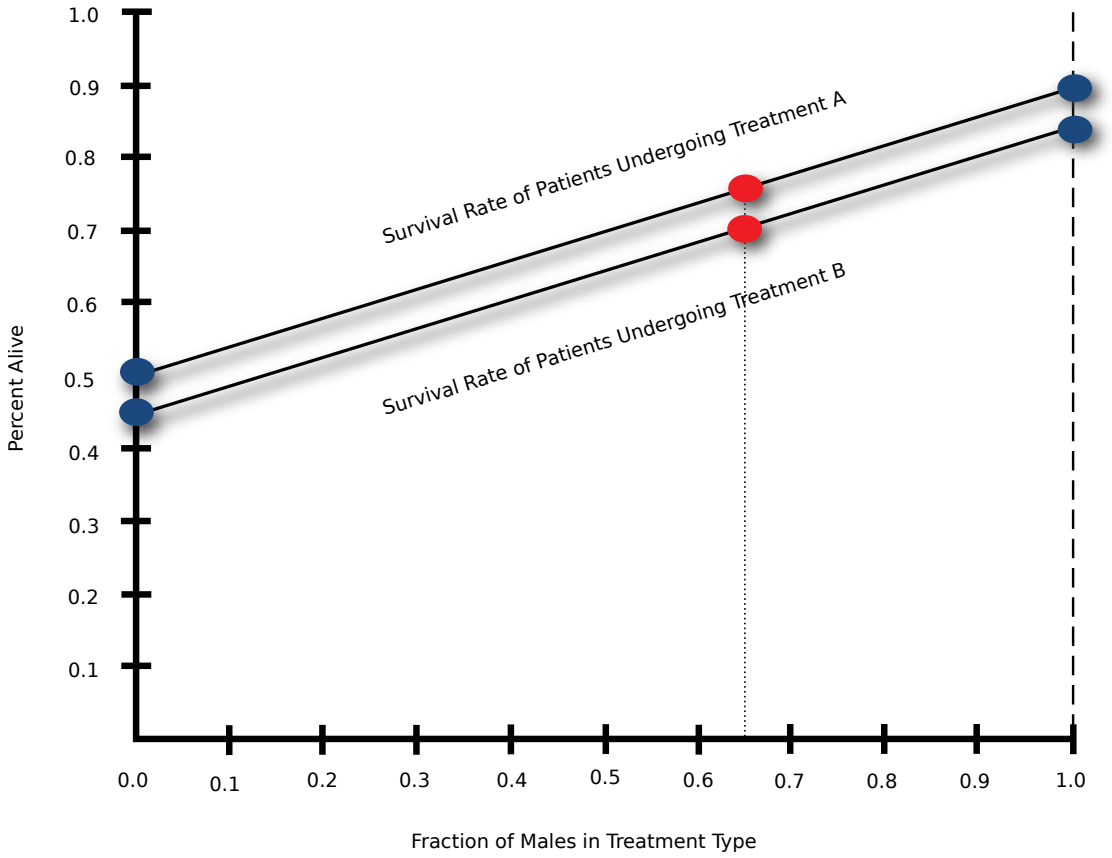

Figure 3: When the grey dots representing the combined population are right above each other in a BK-Plot, Simpson's Paradox cannot occur.

Table 3: Numerical values in Table 1 represented as cell probabilities

	Male	Female
Treatment	Alive	Alive
A	90%	50%
B	85%	45%

and not the combined group, because information on the combined group can be obtained from the subgroups. It can now be shown that representing the numerical values as cell probabilities effectively masks the association between the confounding variable and the independent variable. Consider any change in the number of people in a subgroup of the independent variable (for example: treatment A/B). Note that it would not alter the cell probabilities in any manner. That is, any change in the independent variable which may be caused by the confounding variable will not change our results. We have that the causal pathway between the independent variable and the confounding variable does not change our data, masking the association between

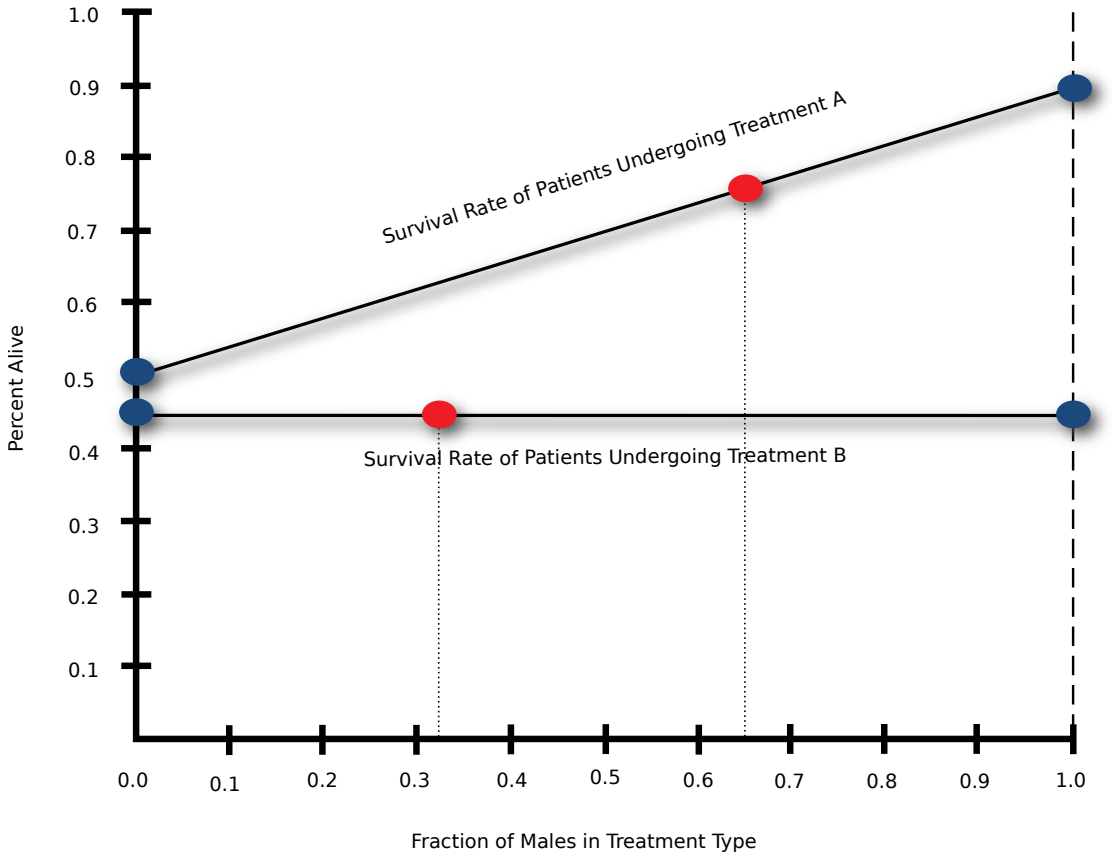

Figure 4: When either of the two line segments in a BK-Plot is parallel to the x-axis, Simpson's Paradox cannot occur.

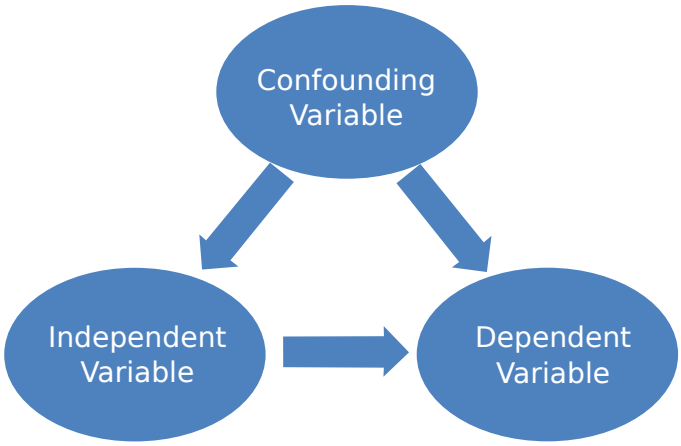

Figure 5: Causal diagram of variables in an occurrence of Simpson's Paradox

the confounding variable and the independent variable.

On the other hand, the association between the independent and dependent variable and the association between the confounding and dependent variable are still preserved. By computing the probability to occur of each subgroup of the confounding variable and by knowing the cell probabilities (Table 4), we can then create a method for making the optimal decision whenever Simpson's Paradox occurs (Figure 6). We can compute adjusted survival rates for treatments A and B by giving the male and female subgroups different weightings. This allows us to eliminate the bias due to the confounding variable, thereby "resolving" the paradox.

Table 4: Table 3 with probability of each gender occurring.

Probability of subject being male: $\frac{360}{900} = (40\%)$

Probability of subject being female: $\frac{540}{900} = (60\%)$

	Male	Female
Treatment	Alive	Alive
A	90%	50%
B	85%	45%

A Real-World Application

Most people today know what cancer is, but few people know about an even deadlier and more common killer: nosocomial infections (NI). Nosocomial infections, or healthcare associated infections, are "hospital" infections obtained through healthcare procedures. Every year, around 1.7 million people contract nosocomial infections, causing almost 100,000 deaths annually in the United States alone. Simpson's Paradox plays an important role in our understanding of nosocomial infections and the development of prevention techniques.

The most frequent type of nosocomial infections in the US are urinary tract infections (UTI), accounting for approximately 40% of all nosocomial infections. Currently, there is no way to guarantee the prevention of nosocomial urinary tract infections, although antibiotic prophylaxis (AB-Proph) is often administered to patients susceptible to nosocomial urinary tract infections. Antibiotic prophylaxis is the use of antibiotics to prevent infection and its complications. In recent years, the use of antibiotic prophylaxis in preventing nosocomial UTI has been under question, as evidence is being produced that point against its effectiveness. While antibiotic prophylaxis has been shown to be effective for preventing bacterial urinary tract infections in some randomized clinical trials, observational studies have also shown antibiotic prophylaxis to be positively associated with urinary tract infection.

In 1997, a large prospective cohort study of nosocomial infections took place in eight hospitals in the Netherlands [11]. A total of 3519 cases of urinary tract infection

were reported in the study, providing valuable information on bacterial urinary tract infections and nosocomial infections in general. The study also contained information regarding the use of antibiotic prophylaxis. Taking a closer look at the data on nosocomial urinary tract infection and administration of antibiotic prophylaxis reveals the presence of Simpson's Paradox. The data was stratified by the incidence rate of UTI in the hospitals that the patients were in and is summarized in the following figure (Figure 6):

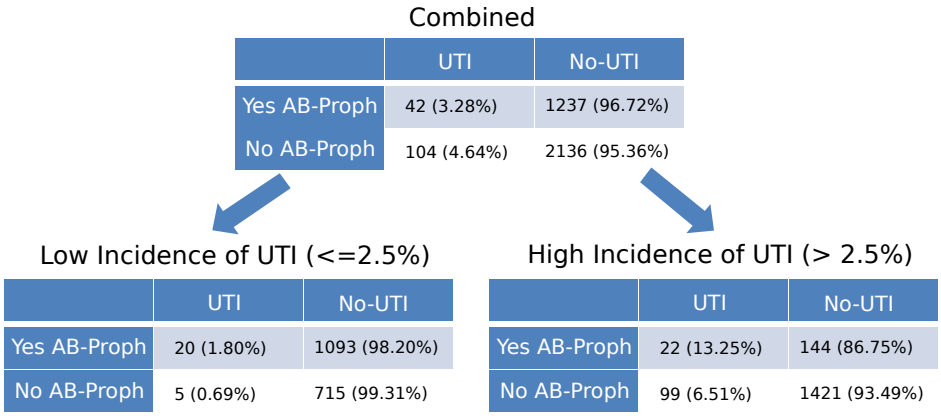

Figure 6: Study data on urinary tract infections

Given the data, we now ask ourselves: "Is an antibiotic prophylaxis effective in preventing nosocomial bacterial urinary tract infection?"

Applying our understanding of Simpson's Paradox, we know that the phenomenon occurs in the data because hospital incidence rate of UTI acts as a confounding variable. The difference in UTI rates in hospitals with low incidence and hospitals with high incidence may be the result of differences in hygiene standards, skill of the doctors, and similar standards. As can be seen from the data, hospitals with high incidence rate of urinary tract infection tend to administer AB-Proph less frequently to patients. Like before, we can resolve the issue by computing the probabilities with weights. For example, the group of patients in hospitals with low incidence of UTI accounts for 52.09% of the population, so we can compute the adjusted UTI incidence rate for the entire sample receiving AB-Proph as

$$0.5209 \times 0.018 + 0.4791 \times 0.1325 = 0.0729.$$

Similarly, the adjusted UTI incidence rate for the entire sample not receiving AB-Proph can be computed as

$$0.5209 \times 0.0069 + 0.4791 \times 0.0651 = 0.0348.$$

The adjusted UTI incidence rates account for the confounding that is present due to Simpson's Paradox. As a result, the values accurately reflect the association between antibiotic prophylaxis use and incidence of UTI. Given that $0.0348 < 0.0729$, under

the assumption that no external variables affect the association between the variables, we can then conclude that the chances of contracting a UTI is smaller for a person of unknown gender receiving no AB-Proph.

It is important to note that in the real world several other factors do exist and could affect the results. However, from a mathematical point of view, we have successfully resolved Simpson's Paradox. Despite the assumptions behind our analysis of Simpson's Paradox, it can be seen that it is a valid approach to understanding the paradox. The insight gained could be used to help prevent millions of cases of UTI every year, along with help resolve other cases of Simpson's Paradox. For the interested reader, a possible extension of Simpson's Paradox involves determining how experimental setups should be designed in order to minimize the probability of Simpson's Paradox from occurring.

Acknowledgements

I would like to express my most sincere thanks to Dr. John C. Howe and Bowei Liu for their guidance and support as my research mentors. I would also like to thank Dr. Howe for his help in reviewing this paper. Special thanks to Dr. Stuart G. Baker for providing research advice in the field.

References

[1] Pearson, K. and Lee, A. and Bramley-Moore, L. "Genetic (reproductive) selection: Inheritance of fertility in man" *Philosophical Translations of the Royal Statistical Society Series A* **173** pp.534–539 (1899)

[2] Yule, G.U. "Notes on the theory of association of attributes in statistics" *Biometrika* **2** pp.121–134 (1903)

[3] Simpson, E.H. "The interpretation of interaction in contingency tables" *Journal of the Royal Statistical Society Series B* **18**:13 pp.238–241 (1951)

[4] Freedman, D.A. "Ecological inference and the ocological fallacy" In Smelser N. and Baltes P. (Eds.) *International Encyclopedia of the Social and Behavioral Sciences* Oxford, UK: Elsevier Science **6** pp.4027–4080 (1999)

[5] Charig, C.R. and Webb, D.R. and Payne, S.R. and Wickham, J.E. "Comparison of treatment of renal calculi by open surgery, percutaneous nephrolithotomy, and extracorporeal shockwave lithotripsy" *Br Med J (C lin R es Ed)* **292**:6524 pp.879–882 (1986)

[6] Gatling, W. and Mulee, M. and Hill, R. "General characteristics of a community-based diabetic population" *Practical Diabetes International* **5**:3 pp.104–107 (1988)

[7] Moore, D.S. and McCabe, G.P. and Craig, B.A. *Introduction to the Practice of Statistics* 6th ed. New York, NY: W.H. Freeman and Company (2009)

[8] Baker, S.G. and Kramer, S.K. "Good for women, good for men, bad for people: Simpsons's paradox and the sex specific analysis in observational studies" *Journal of Women's Health and Gender Based Medicine* **10** pp. 867–872 (2001)

[9] Shapiro, S.H. "Collapsing Contingency Tables - A Geometric Approach" *The American Statistician* **36**:1 pp. 43–36 (1982)

[10] Pavlides, M.G. and Perlman, M.D. "How Likely is Simpson's Paradox?" *American Statistician* **63**:3 pp.226–233 (2009)

[11] Severijnen, A.J., Verbrugh, H.A., Mintjes-de Groot, A.J., Vandenbroucke-Grauls, C.M. and van Pelt, W. "Sentinel System for Nosocomial Infections in the Netherlands: A Pilot Study" *Infect Control Hosp Epidemiol* **18** pp.818–824 (1997)