Simpson's paradox ... and how to avoid it

jojoo64/iStock/Thinkstock

When data from two or more groups are combined, patterns previously seen in the data can reverse or disappear altogether. **H. James Norton** and **George Divine** explain why this happens, and how to prevent it from leading researchers astray

Paradoxes, by their very nature, are intriguing. Those of an inquisitive mind could surely spend hours trying to unpick counter-intuitive results, or seemingly contradictory statements that are, in fact, true.

But paradoxes, especially those of a statistical nature, can also be problematic. Consider a healthcare professional trying to decide whether to prescribe antibiotics in order to reduce the rates of urinary tract infection (UTI) in hospital patients.

Table 1 summarises data from eight hospitals in the Netherlands for patients who do and do not receive prophylactic antibiotics (PAB). The PAB patients have a lower rate of UTI compared to patients without such treatment (3.3% versus 4.6%).[1] It would appear that PAB should be used to prevent UTI.

A cautious healthcare professional, however, might ask a biostatistician to examine the data further. They note that the rate of UTI varies greatly among the eight hospitals from which data has been collected. They also find that PAB use differs among the hospitals.

Table 1. Rate of urinary tract infection by antibiotic prophylaxis

	UTI	*No UTI*	*% with UTI*
Antibiotic prophylaxis	42	1237	**3.3**
No antibiotic prophylaxis	104	2136	**4.6**

Table 2. Urinary tract infection rate by low-incidence hospitals versus high-incidence hospitals

	UTI	*No UTI*	*% with UTI*
Low-incidence hospitals			
Antibiotic prophylaxis	20	1093	**1.8**
No antibiotic prophylaxis	5	715	**0.7**
High-incidence hospitals			
Antibiotic prophylaxis	22	144	**13.3**
No antibiotic prophylaxis	99	1421	**6.5**

Simpson's paradox was described by Udny Yule in 1903 using the hypothetical example of an anti-toxin which could appear to be a "cure" due to a sex-related difference in mortality rates

The biostatistician stratifies the hospitals into two groups – low-incidence hospitals (LIH) and high-incidence hospitals (HIH) – depending on whether the UTI rate is less than or greater than 2.5%. They report the data separately, as in Table 2. This shows that PAB use is associated with an increased rate of UTI in both groups, and that its use is actually detrimental to patients.

Our healthcare professional finds it disconcerting that her present conclusion, that PAB should not be used, is the reverse of what she first thought. "This", the biostatistician explains, "is an example of Simpson's paradox."

Early examples

Simpson's paradox was described by Udny Yule in 1903 using the hypothetical example of a possibly ineffective new anti-toxin which could appear to be a "cure" due to a sex-related difference in mortality rates.[2] Cohen and Nagel are credited with reporting the first example of Simpson's paradox using actual data in 1934.[3] They compare death rates due to tuberculosis between New York City and Richmond, Virginia.

In 1951 Edward Simpson described a fictional example to demonstrate how combining contingency tables may lead to a paradox.[4] In his example, when two tables that have a positive association between a treatment and survival are combined, the association disappears. Twenty years later, Colin Blythe was the first author to name the paradox in honour of Simpson.[5] Moreover, in Blythe's definition and example, the association reverses when the two tables are combined.

A 1975 paper by Bickel *et al*. is given credit for bringing wide attention to Simpson's paradox.[6] When data from 85 graduate school departments at the University of California at Berkeley were aggregated, it was seen that 44% of male applicants were admitted while only 35% of women applicants were accepted. However, when the authors stratified the data by department, this difference disappeared. In fact, the authors concluded that there was a slight bias in favour of the admission of women.

The Berkeley example is not a pure example of Simpson's paradox, however, as a substantial number of the departments went in either direction regarding gender admission rates.

A more robust demonstration of the paradox comes from the field of law and concerns the influence of race on death sentences in the US. One paper showed the death sentence rate versus race of the offender, stratified by race of the victim, for a number of states.[7] The tables for the state of Indiana reveal Simpson's paradox (Table 3). In Indiana whites are nearly twice as likely to receive the death penalty as African-Americans. However, when the data are stratified by the race of the victim, it is African-Americans who have the higher death sentence rate. This occurs both when

Table 3. Indiana death sentence rates by race of murderer and race of victim

		Sentence	
	Jail time	*Death sentences*	*Death sentences/1000*
Combined			
Black offender	2498	28	**11.1**
White offender	2323	49	**20.7**
Black victim			
Black offender	2139	12	**5.6**
White offender	100	0	**0.0**
White victim			
Black offender	359	16	**42.7**
White offender	2223	49	**21.6**

Do's and don'ts

Simpson's paradox reminds us of the philosophical question, "If a tree falls in the forest, and no one is around to hear it, has it made a sound?"

A frequent request for a research statistician is that a customer/researcher emails you a 2×2 table and asks you to supply a p-value (preferably one with $p < 0.05$). You perform the analysis and show an association between the two variables. However, if you had the complete data set and adjusted for a confounding variable and the association went away or reversed, has Simpson's paradox occurred, even if no one knows it? What is the statistician to do to make sure this does not occur?

1. Insist on a statistician being involved in the design, data collection and analysis plan prior to the start of the study. As R. A. Fisher remarked to the Indian Statistical Congress (*c.* 1938): "To call in the statistician after the experiment is done may be no more than asking him to perform a postmortem examination: he may be able to say what the experiment died of."
2. Always think critically about data, especially data from retrospective or observational studies.
3. Discuss with the customer what potential confounding variables are in this particular study.
4. If the variable is on the causal pathway, it is not a confounding variable and you should not adjust for it.
5. Perform statistical analyses to check for confounding.
   (a) If the potential confounding variable is binary, generate stratified 2×2 tables, and compare their results with those in the combined table. If the separate tables have the opposite association compared to the combined table, or if there is an association in the separate tables that is not there when the data are combined (or vice versa), this is evidence of confounding.
   (b) Check for significant associations between the potential confounder and both the outcome and the factor of interest. If both associations exist, that is evidence of confounding.
   (c) Assess the association of interest using logistic regression models: one with just the factor of interest as a predictor, and another with the potential confounder included as a covariate. A rule of thumb states that if the odds ratio estimate for the factor of interest differs between the two models by 10% or more, you should conclude that there is confounding.
6. Make "the sensible interpretation".

the victim is white and when the victim is African-American.

Confounding factors

Simpson's paradox is a manifestation of confounding, and among the dictionary definitions of "confound" are the words "mix up" and "confuse". In the context here, when the relationship between groups and an outcome variable is distorted because of a background factor,[8] that background factor is called a *confounder*. Two of the main conditions needed for a background factor to be a confounder are:

1. The groups differ on the background factor.
2. The background factor influences the outcome variable.

In the example of death sentences, the outcome is a death sentence, the groups are offenders, and the background factor is the race of the victim. For the African-American offenders 85.2% (2151/2526) of the victims are African-American, while for the white offenders 4.2% (100/2372) of the victims are African-American. The murder of an African-American victim results in 0.5% (12/2251) of the perpetrators receiving the death sentence, while in the case of white victims 2.5% (65/2647) of the perpetrators receive the death sentence. Thus, the two main conditions for confounding are met.

We have seen in both the UTI and death sentence examples that stratification of data, instead of combination, may allow a more useful perspective on the data to emerge. However, instead of reporting rates within two or more categories, it can be desirable to generate a single representation of the relationship of interest with the confounding effect removed. This is sometimes accomplished through use of something like the Cochran–Mantel–Haenszel method, or through modelling.

Standardisation

Another method that can be used is standardisation. Standardisation is used in demography to make "fair" comparisons of measures such as death rates or cancer rates between two states or countries that have different age distributions in their populations. A comparison of death rates between Utah and Florida without age adjustment would not be appropriate because Utah has the lowest median age of any of the states, while Florida has one of the highest. We will use the example of UTI from Tables 1 and 2 to show how standardisation can be used to elucidate the phenomenon of Simpson's paradox.

From our first example, recall that the combined overall UTI rate for the eight Dutch hospitals is 3.3% for the PAB group versus 4.6% for the no-PAB group (Table 1). This is not a fair comparison, though, as 87.0% of the PAB group's patients are from the low-incidence group, compared to only 32.1% of the no-PAB group.

Stratification of data, instead of combination, may allow a more useful perspective on the data to emerge. Another method that can be used is standardisation

Let us standardise the rates to make a fair comparison. The standard population will consist of all the patients from both hospitals. The combined population consists of 1833 patients from the LIH group and 1686 patients from the HIH group (52.1% and 47.9%; Table 2). Applying some basic mathematics (see box on page 43), we see that the standardised UTI rate for the PAB group is 7.3%, while the standardised UTI rate for the no-PAB group would be 3.5%. Thus the no-PAB group has a lower UTI rate than the PAB group, which is the opposite result to that seen in the original combined population.

Simpson's second paradox

Simpson also described a second paradox in his paper. That paradox is the following: whether "the sensible interpretation" exists in the separate tables, or is instead found in the combined table, depends upon the context

Table 4. Success of treatment for septic shock by diastolic blood pressure

	Alive	*Dead*	*% alive*
Combined			
Treatment A	860	140	**86**
Treatment B	700	300	**70**
DBP <50			
Treatment A	50	50	**50**
Treatment B	250	250	**50**
DBP ≥50			
Treatment A	810	90	**90**
Treatment B	450	50	**90**

of the data being analysed. This means that the correct interpretation cannot be reliably determined merely by looking at the numbers in the tables.

Suppose (hypothetical) data are analysed to determine whether a new treatment (A) is superior to the standard treatment (B) for septic shock. The combined data show that the proportion surviving to hospital discharge is 86% with treatment A, but only 70% with treatment B. However, if the patients are stratified into two subgroups, depending on whether their diastolic blood pressure (DBP) is less than 50 mmHg, within each stratum (Table 4) the proportions of patients alive at hospital discharge are identical for each treatment.

In all of the previous examples presented, a sensible interpretation has been found in the separate tables. But could it be that the separate tables are not showing the complete story for this situation?

We intentionally omitted crucial details about this hypothetical experiment and the data. Upon arrival to an emergency department, a series of 2000 patients thought to have septic shock were randomised to two equally sized groups of 1000 each, and given treatment A or B, respectively. Contrary to what is implied by Table 4, the two groups of patients actually had identical distributions for their DBP upon arrival.

In this example, all patients survive the first day, and at the end of 24 hours of treatment their DBPs were measured and the DBP categorisation in the table is based upon this second measurement. Does this added information change the sensible interpretation of the data? Only one tenth (100/1000) of patients on treatment A had their DBP crash to below 50, while for treatment B one half (500/1000) of the patients had such a crash. The biology of the situation would suggest that the sensible interpretation is in the combined table. In this case, unlike our other examples, the factor shown in Table 4, DBP, was not a variable that was fixed at the start of the experiment, but was instead an intermediate outcome affected by the treatment.

This illustrates that additional conditions can be required for a factor to be a confounder leading to Simpson's (first) paradox. The condition in this case is that the factor cannot be a consequence of membership in the groups of interest. When this condition is absent, Simpson's second paradox can occur.

In conclusion

Simpson's paradox in its simplest form refers to the reversal of the direction of an association when data from two or more groups are combined to form a single group. The paradox occurs because of the presence of a confounding variable.

It is important to check for confounding variables, as a lack of adjustment for such variables can lead to the wrong conclusion for important issues in medicine, the law, and other fields of study. Stratification and standardisation are two of the statistical techniques that can adjust for confounding.

The standardisation calculations

The standardised UTI rate for the PAB group, assuming a patient population of 52.1% LIH and 47.9% HIH, would be (PAB UTI rate for LIH patients) × (% patients who are LIH in standard population) + (PAB UTI rate for HIH patients) × (% patients who are HIH in standard population).

Therefore, the standardised UTI rate for the PAB group = 0.018 × 0.521 + 0.133 × 0.479 = 0.073 = 7.3%. In a similar manner, the standardised UTI rate for the no-PAB group would be 0.007 × 0.521 + 0.065 × 0.479 = 0.035 = 3.5%.

Acknowledgements

The authors thank Edward Simpson for his suggestions that have improved this article. The reader may be interested in an interview of Edward Simpson that discusses his work as a code-breaker during World War II (*Significance,* June 2010). Simpson was elected a Fellow of the Royal Statistical Society in 1946, but is still only joint ninth in length of service.

References

1. Reintjes, R., de Boer, A., van Pelt, W. and Mintjes-de Groot, J. (2000) Simpson's paradox: An example from hospital epidemiology. *Epidemiology*, **11**(1), 81–83.
2. Yule, G. U. (1903) Notes on the theory of association of attributes in Statistics. *Biometrika*, **2**(2), 121–134.
3. Cohen, M. R. and Nagel, E. (1934) *An Introduction to Logic and Scientific Method*. New York: Harcourt, Brace.
4. Simpson, E. H. (1951) The interpretation of interaction in contingency tables. *Journal of the Royal Statistical Society*, *Series B*, **13**, 238–241.
5. Blyth, C. R. (1972) On Simpson's paradox and the sure thing principle. *Journal of the American Statistical Association*, **67**, 364–366.
6. Bickel, P. J., Hammel, E. A. and O'Connell, J. W. (1975) Sex bias in graduate admissions: Data from Berkeley. *Science*, **187**, 398–404.
7. Blume, J. H., Eisenberg, T. and Wells, M. T. (2004) Explaining death row's population and racial composition. *Journal of Empirical Legal Studies*, **1**(1), 165–207.
8. Anderson, S., Auquier, A., Hauck, W. W., Oakes, D., Vandaele, W. and Weisberg, H. I. (1980) *Statistical Methods for Comparative Studies: Techniques for Bias Reduction*. New York: John Wiley & Sons.

H. James (Jim) Norton is director of biostatistics at Dickson Advanced Analytics, Carolinas Medical Center, Charlotte, North Carolina. He taught undergraduate biostatistics for 20 years, mostly at the University of North Carolina at Charlotte

George Divine is a senior research biostatistician at Henry Ford Hospital, Detroit, Michigan. At the Joint Statistical Meetings, Norton and Divine have won the "Best Contributed Paper" twice in the Section on Statistical Education and three times in the Section on Teaching Statistics in the Health Sciences