

SIMPSON'S PARADOX

Palak Mishra
Y21 SDS
210690

INTRODUCTION

Simpson's paradox is a phenomenon in probability and statistics in which a trend appears in several groups of data but disappears or reverses when the groups are combined.

Simpson's paradox is important for three critical reasons. First, people often expect statistical relationships to be immutable. They often are not. The relationship between two variables might increase, decrease, or even change direction depending on the set of variables being controlled. Second, Simpson's paradox is not simply an obscure phenomenon of interest only to a small group of statisticians. Simpson's paradox is actually one of a large class of association paradoxes. Third, Simpson's paradox reminds researchers that causal inferences, particularly in non-experimental studies, can be hazardous. Uncontrolled and even unobserved variables that would eliminate or reverse the association observed between two variables might exist.

HISTORY

Edward H. Simpson was first to describe this phenomenon in a technical paper in 1951, hence, the name. Meanwhile, the name 'Simpson's paradox' was introduced by **Colin R. Blyth** in 1972. Simpson's Paradox is also referred to as Simpson's reversal, the Yule–Simpson effect, the amalgamation paradox, or the reversal paradox.

REAL-LIFE EXAMPLES

One of the best-known examples of Simpson's paradox comes from a study of **gender bias** among graduate school admissions to University of California, Berkeley. The admission figures for the fall of 1973 showed that men applying were more likely than women to be admitted, and the difference was so large that it was unlikely to be due to chance. However, when taking into account the information about departments being applied to, the different rejection percentages reveal the different difficulty of getting into the department, and at the same time it showed that women tended to apply to more competitive departments with lower rates of admission, even among qualified applicants (such as in the English department), whereas men tended to apply to less competitive departments with higher rates of admission (such as in the engineering department). The pooled and corrected data showed a "small but statistically significant bias in favor of women".

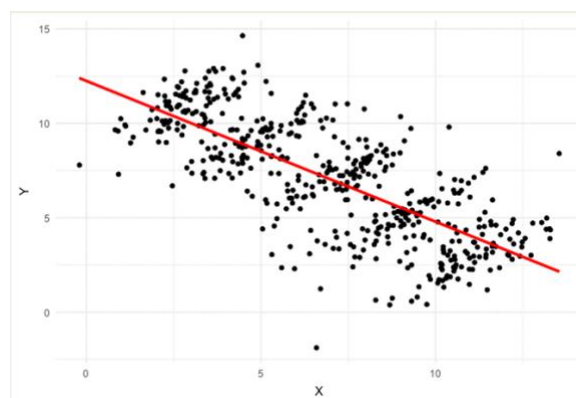
Another example comes from a real-life medical study comparing the success rates of **two treatments for kidney stones**. Treatment A includes open surgical procedures and Treatment B includes closed surgical procedures. The paradoxical conclusion is that treatment A is more effective when used on small stones, and also when used on large stones, yet treatment B appears to be more effective when considering both sizes at the same time. In this example, the factor causing the paradox is the size of the stones, which was not previously known to researchers to be important until its effects were included.

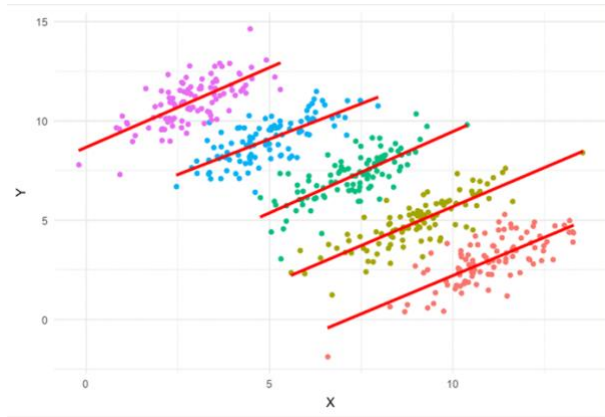
A common example of Simpson's paradox involves the batting averages of players in professional baseball. It is possible for one player to have a higher batting average than another player each year for a number of years, but to have a lower batting average across all of those years. This phenomenon can occur when there are large differences in the number of at-bats between the years.

CAUSES

Simpson's Paradox can occur for various reasons, including:

1. **Confounding variables:** Hidden factors that affect both the independent and dependent variables, leading to a misleading overall trend.
2. **Sample size differences:** When subgroups have significantly different sample sizes, the aggregation of data may mask the underlying trends in individual subgroups.
3. **Measurement error:** Inaccuracies in data collection can contribute to the paradox.
4. **Heterogeneity:** Subgroups may have different characteristics or properties that affect the overall trend when combined.
5. **Interaction effects:** Variables can interact in complex ways, leading to paradoxical results when considered together.





SOLUTION

Avoiding Simpson's Paradox requires careful and thoughtful data analysis. Here are some strategies to help prevent or mitigate its occurrence:

1. **Understand your data:** Thoroughly understand the characteristics of your dataset, including potential confounding variables and subgroups.
2. **Explore subgroups:** Examine the data within subgroups before drawing conclusions from the overall dataset. This can reveal any paradoxical trends that might be hidden when aggregated.
3. **Control for confounders:** When possible, control for confounding variables through statistical techniques such as regression analysis or matching. This helps isolate the true relationships between variables.
4. **Consider causality:** Be cautious about inferring causality based solely on observed associations. Simpson's Paradox often highlights the importance of considering causality more carefully.
5. **Increase sample size:** If possible, increase the sample size in subgroups to reduce the impact of sample size differences.
6. **Sensitivity analysis:** Perform sensitivity analyses to assess how results change when different assumptions or modeling choices are made. This can help you understand the robustness of your findings.
7. **Visualize the data:** Use data visualization techniques to help identify and understand trends within subgroups and the overall dataset. Graphs and charts can make hidden patterns more apparent.
8. **Consult experts:** If you're unsure about potential sources of Simpson's Paradox in your data, consult with statisticians or domain experts who can provide valuable insights.

CONCLUSION

Remember that Simpson's Paradox is a reminder that data analysis is complex, and drawing conclusions from aggregate data without considering underlying factors can lead to misleading results. Careful data exploration and analysis are key to avoiding or addressing this paradox.