Countering Simpson's Paradox with Counterfactuals

Arran Zeyu Wang*
UNC-Chapel Hill

David Borland†
RENCI, UNC-Chapel Hill

David Gotz‡
UNC-Chapel Hill

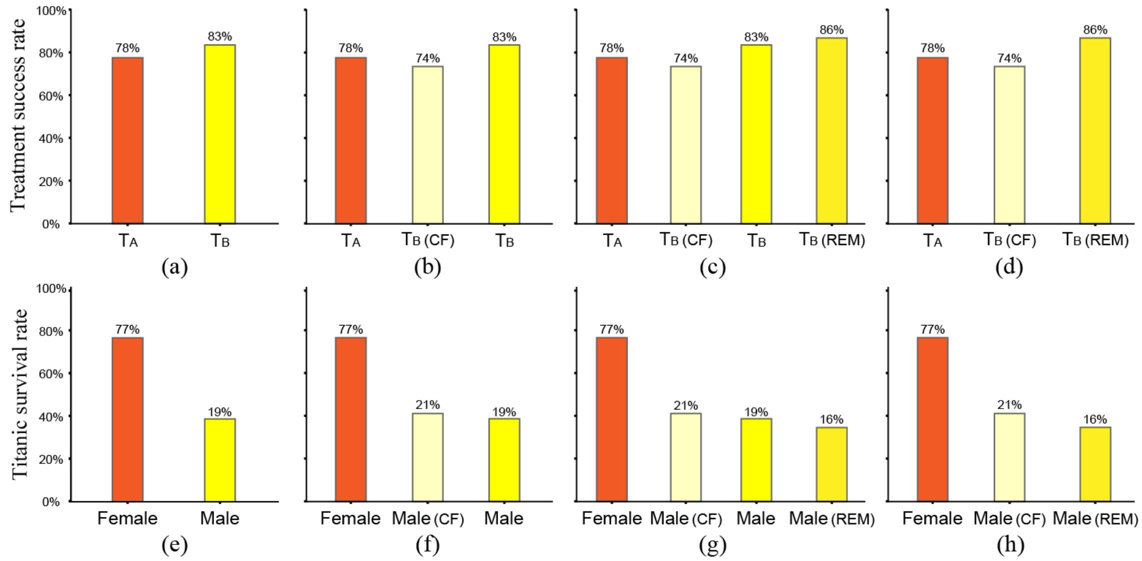

Figure 1: The first row shows four visualizations of the kidney stone treatment dataset [5] from Table 1: (a) a traditional visualization of treatment A (T_A) and treatment B (T_B), (b)-(d) alternative visualization designs that include the counterfactual subset ($T_{B(CF)}$) and remainder subset ($T_{B(REM)}$). The second row shows four visualizations of the Titanic survival dataset [1] from Table 2: (e) a traditional visualization of female and male passengers, (f)-(h) alternative visualizations including the counterfactual subset ($Male_{(CF)}$) and remainder subset ($Male_{(REM)}$).

Abstract

Visualizations are widely used to compare aggregate statistics between subsets of data. However, aggregation can often obscure patterns or trends and produce misleading views of the data. One example of this risk is Simpson's Paradox, a phenomenon that can commonly occur in interactive data visualizations that enable ad hoc grouping and filtering. We explore the potential of counterfactuals—widely used in causal inference— to help counter the risks of invalid conclusions due to Simpson's Paradox in data visualization.

1 Introduction

Aggregation is widely used in visualization to show summary statistics for groups of data points. Providing interactive capabilities to navigate different levels of aggregation and apply filters can enable exploratory analysis. These capabilities, although powerful, can also introduce risks. One such risk is Simpson's Paradox, a phenomenon in which trends that appear at one level of aggregation may disappear or reverse when data is subdivided into lower levels of aggregation.

For example, one widely-cited real-world example comes from an analysis of alternative medical treatments for kidney stones [5]. As shown in Table 1, the study included patients with stones of variable size, classified as large or small. Compared to Treatment B (T_B), Treatment A (T_A) performed best on small stones and best on large stones. However, counter-intuitively, Treatment B appeared to have a higher success rate overall. This paradox is a result of the unequal distribution of large and small stone patients assigned to each treatment.

During exploratory analysis, this type of reversal of trends at different levels of aggregation can happen without user awareness and can lead users to make incorrect conclusions. This motivates research exploring ways to counter such problems, including the work exploring the use of counterfactuals described here.

*e-mail: zeyuwang@cs.unc.edu
†e-mail: borland@renci.org
‡e-mail: gotz@unc.edu

2 Related Work

Two broad areas of prior research inform the proposed use of counterfactuals for countering Simpson's paradox. First, prior research has explored visual ways to communicate or mitigate Simpson's paradox and related phenomena. This includes visualizations that show data concurrently at multiple levels of aggregation (e.g., [2]) to facilitate comparisons. In our own work, we have explored visualizations of selection bias to identify when subgroups have potentially important differences [3], as well as to adjust samples by applying weights that facilitate more appropriate comparisons [4].

The second area of related work is focused on the concept of counterfactuals. Counterfactual reasoning is a fundamental concept in statistical causal inference [7]. This approach is based on the idea of constructing hypothetical scenarios ("what if things were the same *except* for this one fact?") and then making inferences about what would happen under those counterfactual conditions. In the context of visualization, counterfactuals have been applied to improve model interpretability [8] and support more accurate inferences of causal relationships from visualizations [6].

3 Countering Simpson's Paradox

As explained in the kidney stone example (Sect. 1), the misleading aggregate success rates leading to Simposon's Paradox are due to

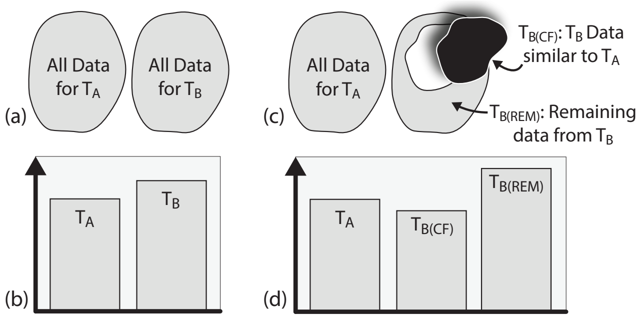

Figure 2: (a-b) A typical visualization comparing two groups T_A and T_B. (c) A counterfactual subset $T_{B(CF)}$ contains data points just like those in T_A. This also creates $T_{B(REM)}$ with the remaining data points from T_B. (d) A comparison of T_A and $T_{B(CF)}$ avoids Simpson's Paradox.

differences in the T_A and T_B populations. The easier-to-treat patients with small stones were much more likely to receive T_B, making the overall success rate for T_B higher even though it was less effective than T_A.

To apply counterfactual reasoning to this problem, one can ask *"What if we had a population of patients exactly like those treated with T_A, except that we treated them with T_B instead?"* While this counterfactual is not represented directly within the data, it can be simulated by sampling from the population receiving T_B a subset of patients similar to those patients receiving T_A. We refer to this sample as the counterfactual subset of T_B, or $T_{B(CF)}$. See Fig. 2.

$T_{B(CF)}$ will comprise a group of patients with similar variable distributions to T_A. In this simple example, the only attribute in addition to those already visualized (treatment type and success) is kidney stone size. We therefore sample a group of patients from T_B with the same ratio of large:small kidney stones as T_A to include in $T_{B(CF)}$. The T_B patients remaining after this sampling process are noted as $T_{B(REM)}$.

Because, by construction, the subset $T_{B(CF)}$ has the same distribution of kidney stone sizes as T_A, aggregate statistics such as the success rate of the treatment can be fairly compared within a visualization. The remaining subgroup of T_B patients, $T_{B(REM)}$, can also be visually compared to $T_{B(CF)}$ to see how the differences in subgroup composition (in this case, differences in stone size) may have impacted the aggregate statistics (in this case, the success rate of Treatment B).

It is important to note that selecting the counterfactual subset is itself a significant challenge. The example outlined here has a straightforward solution due to the fact that there is just a single attribute (stone size) known about individual patients in each treatment group. The approach outlined above can scale to higher-dimensional datasets (a benefit versus prior approaches mentioned in Sect. 2), but the selection of the counterfactual subset becomes a more complex step in the process.

4 Examples

To demonstrate how this counterfactual approach can be applied, we provide two examples using a pair of simple real-world data sets. The first uses data from the kidney stone treatment study introduced in Sect. 1 [5]. The data for this study is shown in Table 1, where treatment T_A outperforms T_B for both small and large stones, but T_B appears to succeed at a higher rate when viewed in aggregate due to different distributions of stone sizes. In this example, $T_{B(CF)}$ is sampled from T_B as shown in the table. The success rate for $T_{B(CF)}$ is worse than T_A, which can be visualized as shown in Fig. 1 to more accurately communicate the desired comparison between treatments.

We note that the remaining patients in $T_{B(REM)}$, as we would expect, have "easier to treat" small stones which were over-represented in the original T_B. If Simpson's Paradox is present, we should expect to see the pattern displayed in Fig. 1(d). More specifically, Fig. 1(d) shows a relatively large difference between the counterfactual (CF) and remaining (REM) subsets of T_B, with the difference straddling the T_A value.

The second example applies the counterfactual approach to survival data from the RMS Titanic [1]. This data, shown in Table 2, describes survival rates for passengers by gender and cabin class. In this case, there is no occurrence of Simpson's Paradox. Females survived at a higher rate overall and across all classes, even though the distribution of cabin class differed by sex. Applying the counterfactual approach in this case results in the data shown in the final two columns of Table 2 and illustrated in Fig. 1(e)-(h). Unlike the kidney stone example, in this case, the CF and REM subgroups show little difference, thus Simpson's paradox is not present (see Fig. 1(h)).

Table 1: Success rates of kidney stone treatments (T_A and T_B) for small and large stones [5]. Values in **bold** imply that T_B is more effective overall (83%), however, T_A is more effective for both small (93%) and large (73%) stones individually, resulting in Simpson's Paradox.

Stone Size	T_A	T_B	$T_{B(CF)}$	$T_{B(REM)}$
Small	**93%** (81/87)	87% (234/270)	88% (23/26)	86% (211/244)
Large	**73%** (192/263)	69% (55/80)	69% (55/80)	N/A (0/0)
All	78% (273/350)	**83%** (289/350)	74% (78/106)	86% (211/244)

Table 2: Survival rates from the RMS Titanic [1] for male and female passengers per cabin class, with highest rates in **bold**. Simpson's Paradox is not present in this case.

Cabin	Female	Male	$Male_{(CF)}$	$Male_{(REM)}$
Class 1	**97%** (91/94)	37% (45/122)	37% (35/94)	36% (10/28)
Class 2	**92%** (70/76)	16% (17/108)	16% (12/76)	16% (5/32)
Class 3	**56%** (81/144)	14% (47/347)	14% (20/144)	13% (27/203)
All	**77%** (242/314)	19% (109/577)	21% (67/314)	16% (42/263)

Acknowledgments

This research is made possible in part by NSF Award #2211845.

References

[1] Titanic survival dataset. https://www.kaggle.com/c/titanic/data.

[2] Z. Armstrong and M. Wattenberg. Visualizing Statistical Mix Effects and Simpson's Paradox. *IEEE Transactions on Visualization and Computer Graphics*, 20(12):2132–2141, Dec. 2014. Conference Name: IEEE Transactions on Visualization and Computer Graphics.

[3] D. Borland, W. Wang, J. Zhang, J. Shrestha, and D. Gotz. Selection Bias Tracking and Detailed Subset Comparison for High-Dimensional Data. *IEEE Transactions on Visualization and Computer Graphics*, 26(1), 2020.

[4] D. Borland, J. Zhang, S. Kaul, and D. Gotz. Selection-Bias-Corrected Visualization via Dynamic Reweighting. *IEEE Transactions on Visualization and Computer Graphics*, 27(2):1481–1491, 2021. Conference Name: IEEE Transactions on Visualization and Computer Graphics.

[5] C. R. Charig, D. R. Webb, S. R. Payne, and J. E. Wickham. Comparison of treatment of renal calculi by open surgery, percutaneous nephrolithotomy, and extracorporeal shockwave lithotripsy. *British Medical Journal (Clinical research ed.)*, 292(6524):879–882, Mar. 1986.

[6] S. Kaul, D. Borland, N. Cao, and D. Gotz. Improving visualization interpretation using counterfactuals. *IEEE Transactions on Visualization and Computer Graphics*, 28(1):998–1008, 2021.

[7] J. Pearl, M. Glymour, and N. Jewell. *Causal Inference in Statistics: A Primer*. Wiley, 2016.

[8] J. Wexler, M. Pushkarna, T. Bolukbasi, M. Wattenberg, F. Viégas, and J. Wilson. The what-if tool: Interactive probing of machine learning models. *IEEE transactions on visualization and computer graphics*, 26(1):56–65, 2019.