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 ($T_B(\text{CF})$) and remainder subset ($T_B(\text{REM})$). 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 ($\text{Male(\text{CF})}$) and remainder subset ($\text{Male(\text{REM})}$).

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.

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 Simpson’s Paradox are due to

---

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

<table>
<thead>
<tr>
<th>Stone Size</th>
<th>( T_A )</th>
<th>( T_B )</th>
<th>( T_B^{(CF)} )</th>
<th>( T_B^{(REM)} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Small</td>
<td>93% (81/87)</td>
<td>87% (234/270)</td>
<td>83% (97/116)</td>
<td>86% (211/244)</td>
</tr>
<tr>
<td>Large</td>
<td>73% (192/263)</td>
<td>76% (558/730)</td>
<td>86% (211/244)</td>
<td>69% (558/80)</td>
</tr>
<tr>
<td>All</td>
<td>78% (273/350)</td>
<td>83% (289/350)</td>
<td>78% (273/350)</td>
<td>86% (211/244)</td>
</tr>
</tbody>
</table>

### 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.

<table>
<thead>
<tr>
<th>Cabin</th>
<th>Female</th>
<th>Male</th>
<th>( T_B^{(CF)} )</th>
<th>( T_B^{(REM)} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Class 1</td>
<td>97% (91/94)</td>
<td>37% (45/122)</td>
<td>37% (35/94)</td>
<td>36% (10/28)</td>
</tr>
<tr>
<td>Class 2</td>
<td>92% (70/76)</td>
<td>16% (17/108)</td>
<td>16% (12/76)</td>
<td>16% (5/32)</td>
</tr>
<tr>
<td>Class 3</td>
<td>56% (81/144)</td>
<td>14% (47/347)</td>
<td>14% (20/144)</td>
<td>13% (27/203)</td>
</tr>
<tr>
<td>All</td>
<td>77% (242/314)</td>
<td>19% (109/577)</td>
<td>21% (67/314)</td>
<td>16% (42/253)</td>
</tr>
</tbody>
</table>

### ACKNOWLEDGMENTS

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

### REFERENCES