Simpson's Paradox

Andy Grogan-Kaylor

17 Aug 2021 10:16:48

Introduction

Simpson's Paradox is the idea that associations between variables that are found between two variables in the sample as a whole, can be *very different* (or even *reversed*) when a third variable is introduced.

There are specific mathematical conditions under which Simpson's Paradox applies. However, thinking through these mathematical conditions may not be helpful to develop intuitions about Simpson's Paradox.

Careful inspection of visualizations can help us to develop some understanding of Simpson's Paradox.

> The key substantive conclusion is that models of the data that only account for two variables may provide *very different*–or even *opposite*–conclusions from models of the world that account for many variables.

Some Hypothetical Data

These data are based on the hypothetical data provided by Simpson (1951).

```
. use simpsonsparadox.dta, clear

. list // list out the data
```

	count	treatment	status	group
1.	4	untreated	alive	A
2.	3	untreated	dead	A
3.	8	treated	alive	A
4.	5	treated	dead	A
5.	2	untreated	alive	B
6.	3	untreated	dead	B
7.	12	treated	alive	B
8.	15	treated	dead	B

> Because these are weighted data–where every row of data represents more than one individual–we will need to use the `[fweight=count]` syntax.

Graph Scheme

I am not a particular fan of Stata's default graphing scheme, so I am going to use the `michigan` graph scheme here. (To install, type `net from https://agrogan1.github.io/Stata`)

```
. set scheme michigan // use Michigan graph scheme
```

Sample As A Whole

The treatment appears to have NO effect. An equal percentage of those receiving and not receiving treatment are alive.

Cross Tabulation

```
. tabulate status treatment [fweight=count], row col
```

Key
frequency *row percentage* *column percentage*

status	treatment untreated	treated	Total
dead	6	20	26
	23.08	76.92	100.00
	50.00	50.00	50.00
alive	6	20	26
	23.08	76.92	100.00
	50.00	50.00	50.00
Total	12	40	52
	23.08	76.92	100.00
	100.00	100.00	100.00

Mosaic Plot

Mosaic Plots are a little bit counterintuive at first. However, I believe that–after a little bit of study–Mosaic Plots provide the best visual representation of these relationships. The command for a Mosaic Plot in Stata is `spineplot`, installed by typing `ssc install spineplot`.

```
. spineplot status treatment [fweight=count], ///
> bar1(color(gold) fintensity(%100)) bar2(color(navy %100) fintensity(%100)) ///
> title("The Treatment Appears To Have NO Effect") ///
> subtitle("In The Sample As A Whole") ///
> note("An Equal % Of Those Receiving And Not Receiving Treatment Are Alive")
(note:  named style % 100 not found in class intensity, default attributes used)
(note:  named style % 100 not found in class intensity, default attributes used)

. graph export my_Simpsons_spineplot.png, width(1000) replace
file /Users/agrogan/Desktop/GitHub/agrogan1.github.io/myposts/my_Simpsons_spineplot.png
    saved as PNG format
```

Bar Chart

It is sometimes said that every data visualization should, or could, be a bar chart, so I present one here. Bar charts are a common and useful type of data visualization.

```
. graph bar [fweight=count], over(status) over(treatment) asyvars ///
> title("The Treatment Appears To Have NO Effect") ///
> subtitle("In The Sample As A Whole") ///
> note("An Equal % Of Those Receiving And Not Receiving Treatment Are Alive")
```

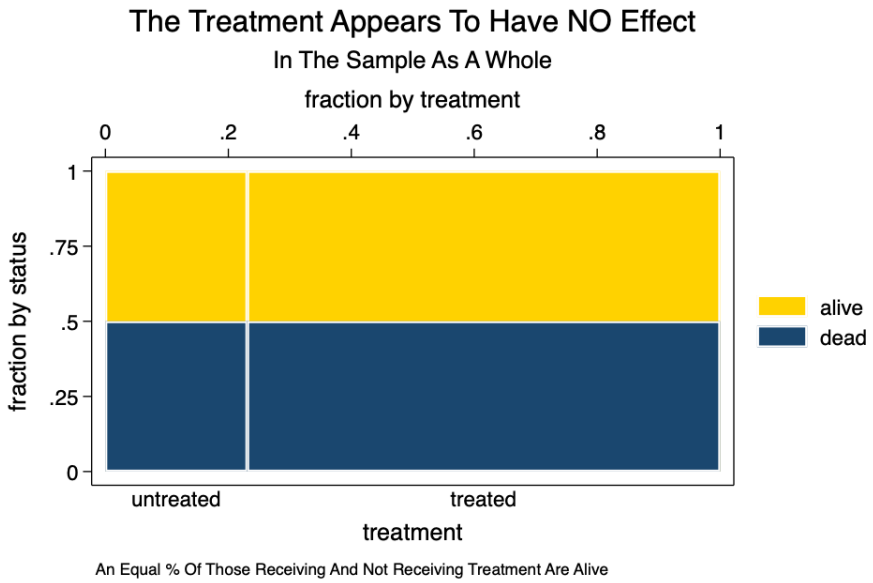

Figure 1: Spineplot or Mosaic Plot

```
. graph export my_Simpsons_barchart.png, width(1000) replace
file /Users/agrogan/Desktop/GitHub/agrogan1.github.io/myposts/my_Simpsons_barchart.png saved
    as PNG format
```

Pie Chart

A Pie Chart sometimes is more intuitive at first, but actually has less information than a Mosaic Chart.

```
. graph pie, over(status) by(treatment, ///
> title("The Treatment Appears To Have NO Effect") ///
> subtitle("In The Sample As A Whole") ///
> note("An Equal % Of Those Receiving And Not Receiving Treatment Are Alive"))

. graph export my_Simpsons_piechart.png, width(1000) replace
file /Users/agrogan/Desktop/GitHub/agrogan1.github.io/myposts/my_Simpsons_piechart.png saved
    as PNG format
```

Sample Divided By Groups

The treatment appears to have an effect. A greater percentage of those recieving treatment are alive.

Cross Tabulation

```
. bysort group: tabulate status treatment [fweight=count], row col
```

```
-> group = A

┌─────┐
│ Key │
```

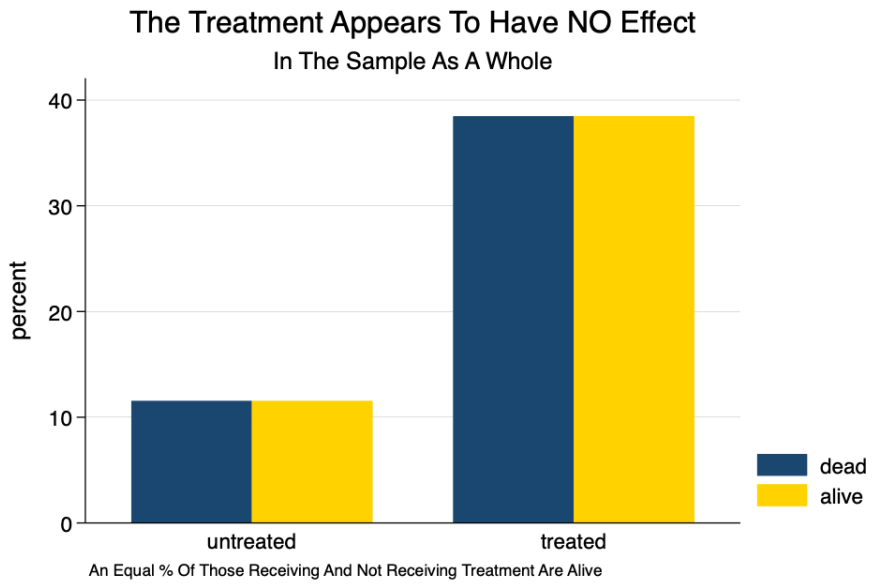

Figure 2: Bar Chart

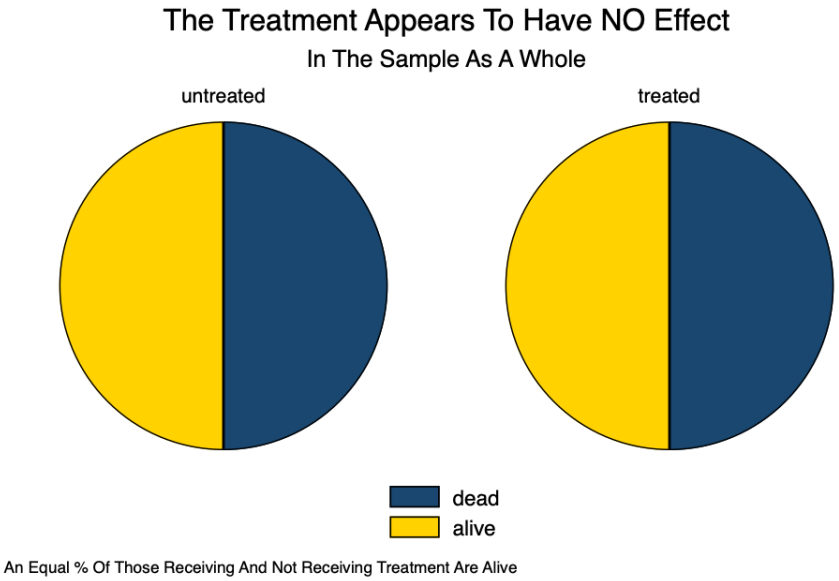

Figure 3: Pie Chart

frequency row percentage column percentage

status	treatment untreated	treated	Total
dead	3	5	8
	37.50	62.50	100.00
	42.86	38.46	40.00
alive	4	8	12
	33.33	66.67	100.00
	57.14	61.54	60.00
Total	7	13	20
	35.00	65.00	100.00
	100.00	100.00	100.00

```
-> group = B
```

Key
frequency row percentage column percentage

status	treatment untreated	treated	Total
dead	3	15	18
	16.67	83.33	100.00
	60.00	55.56	56.25
alive	2	12	14
	14.29	85.71	100.00
	40.00	44.44	43.75
Total	5	27	32
	15.62	84.38	100.00
	100.00	100.00	100.00

Mosaic Plot

```
. spineplot status treatment [fweight=count] if group == 1, ///
> bar1(color(gold) fintensity(%100)) bar2(color(navy) fintensity(%100)) ///
> caption("Group A") ///
> name(spineplotA, replace)
(note:  named style % 100 not found in class intensity, default attributes used)
(note:  named style % 100 not found in class intensity, default attributes used)

. spineplot status treatment [fweight=count] if group == 2, ///
> bar1(color(gold) fintensity(%100)) bar2(color(navy) fintensity(%100)) ///
> caption("Group B") ///
> name(spineplotB, replace)
(note:  named style % 100 not found in class intensity, default attributes used)
(note:  named style % 100 not found in class intensity, default attributes used)

. graph combine spineplotA spineplotB, ///
> title("The Treatment Does Appear To Have An Effect") ///
> subtitle("When Examined By Group") ///
> note("A Greater % Of Those Receiving Treatment Are Alive")
(note:  named style % 100 not found in class intensity, default attributes used)
(note:  named style % 100 not found in class intensity, default attributes used)
(note:  named style % 100 not found in class intensity, default attributes used)
(note:  named style % 100 not found in class intensity, default attributes used)
```

```
. graph export my_Simpsons_spineplot2.png, width(1000) replace
file /Users/agrogan/Desktop/GitHub/agrogan1.github.io/myposts/my_Simpsons_spineplot2.png
    saved as PNG format
```

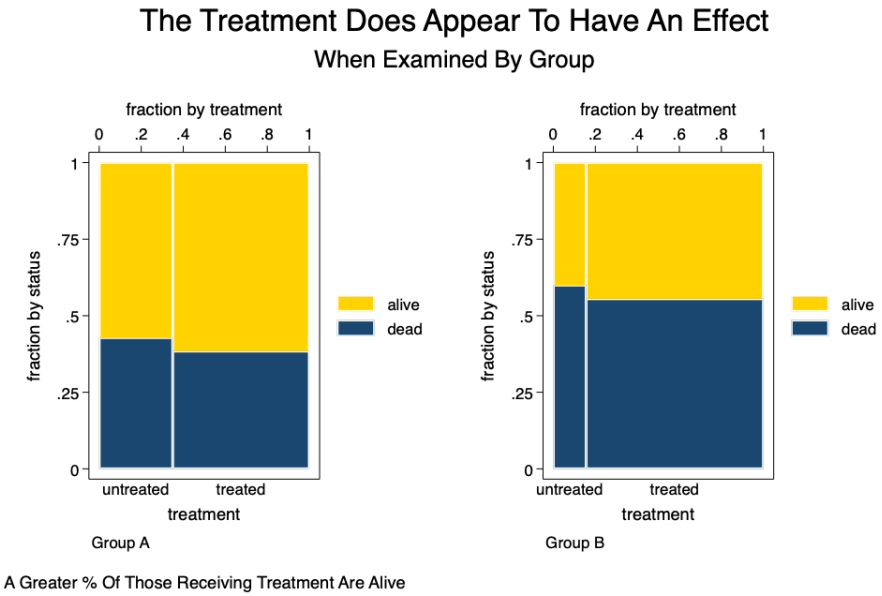

Figure 4: Spineplot or Mosaic Plot

Bar Chart

```
. graph bar [fweight=count], over(status) over(treatment) asyvars ///
> by(group, ///
> title("The Treatment Does Appear To Have An Effect") ///
> subtitle("When Examined By Group") ///
> note("A Greater % Of Those Receiving Treatment Are Alive"))

. graph export my_Simpsons_barchart2.png, width(1000) replace
file /Users/agrogan/Desktop/GitHub/agrogan1.github.io/myposts/my_Simpsons_barchart2.png
    saved as PNG format
```

Pie Chart

```
. graph pie if group == 1 [fweight=count], over(status) by(treatment, ///
> caption("Group A")) ///
> name(piechartA, replace)

. graph pie if group == 2 [fweight=count], over(status) by(treatment, ///
> caption("Group B")) ///
> name(piechartB, replace)

. graph combine piechartA piechartB, ///
> title("The Treatment Does Appear To Have An Effect") ///
> subtitle("When Examined By Group") ///
> note("A Greater % Of Those Receiving Treatment Are Alive")

. graph export my_Simpsons_piechart2.png, width(1000) replace
file /Users/agrogan/Desktop/GitHub/agrogan1.github.io/myposts/my_Simpsons_piechart2.png
    saved as PNG format
```

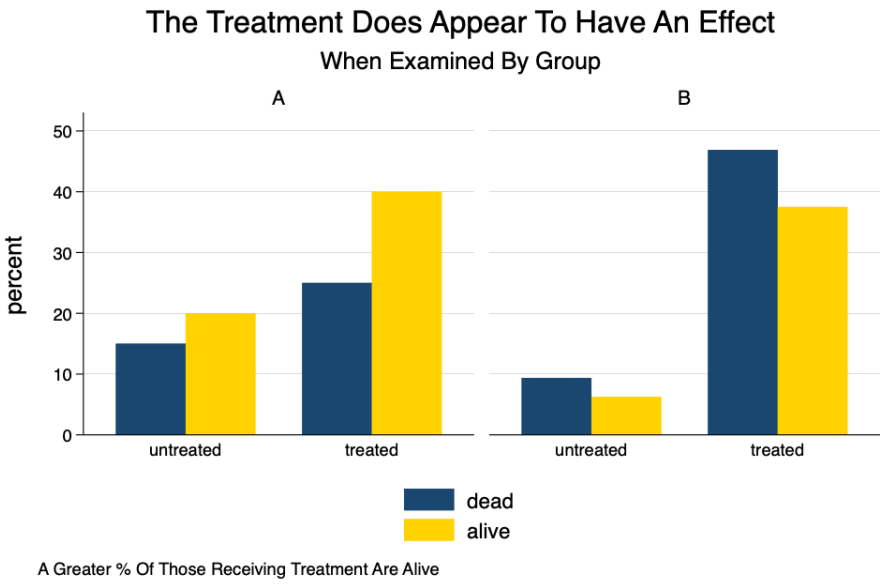

Figure 5: Bar Chart

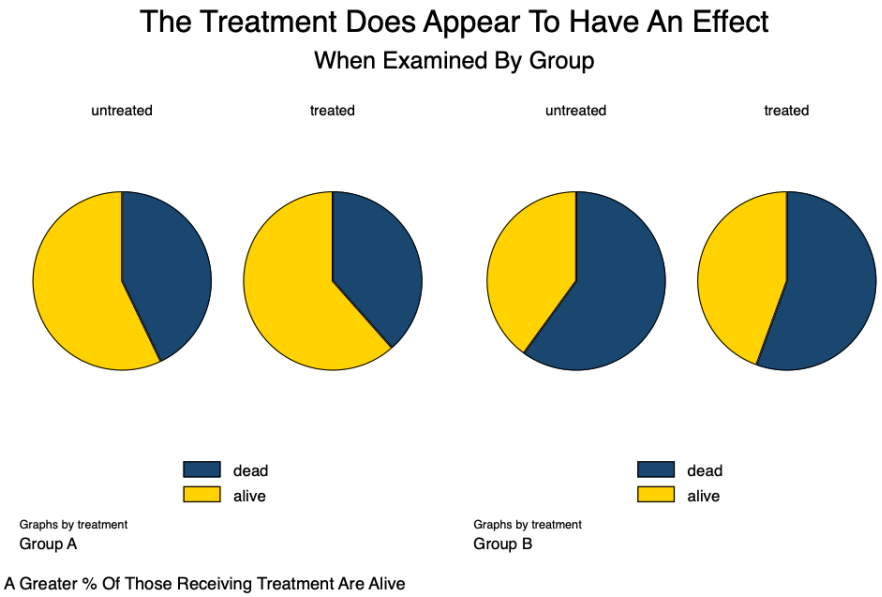

Figure 6: Pie Chart

Reference

Simpson, E. (1951). The Interpretation of Interaction in Contingency Tables. *Journal of the Royal Statistical Society. Series B (Methodological)*, 13(2), 238-241. Retrieved February 2, 2021, from http://www.jstor.org/stable/2984065