Simpson's Paradox

Andy Grogan-Kaylor

2021-03-06

Contents

1 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.

2 Some Hypothetical Data

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

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

3 *Sample As A Whole*

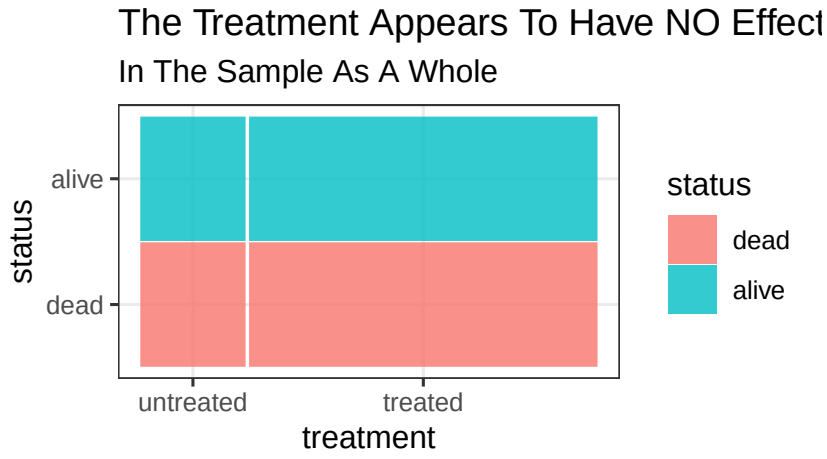
The treatment appears to have no effect.

3.1 *Cross Tabulation*

	dead	alive
untreated	6	6
treated	20	20

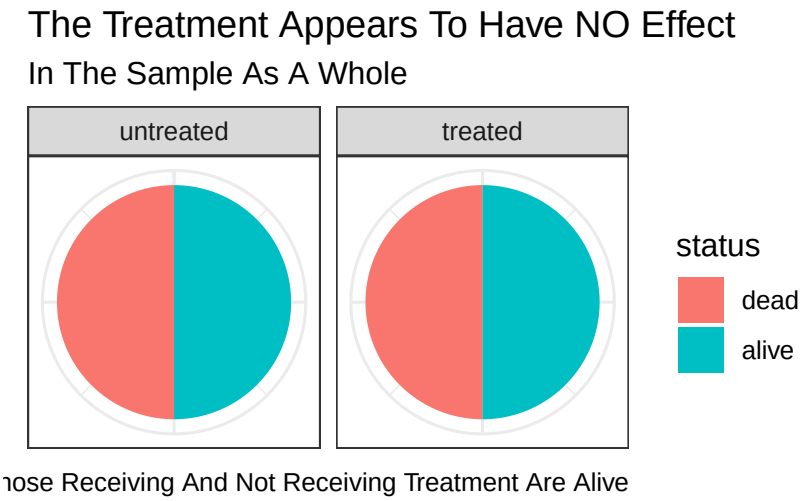
3.2 *Mosaic Plot*

Mosaic Plots are a little bit counterintuive at first. However, I believe they provide the best visual representation of these relationships.



hose Receiving And Not Receiving Treatment Are Alive

3.3 *Pie Chart*



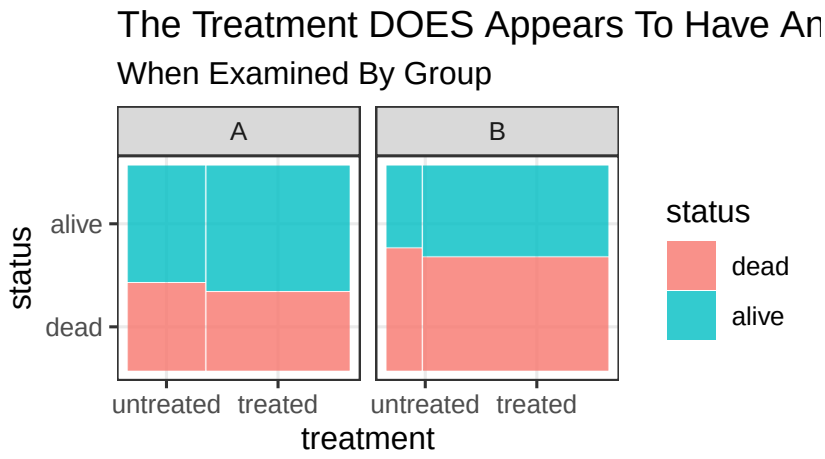
4 *Sample Divided By Groups*

The treatment appears to have an effect.

4.1 *Cross Tabulation*

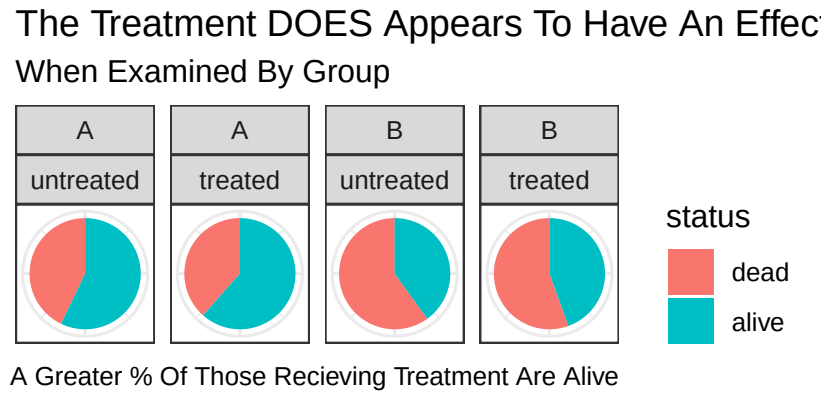
		A	B
untreated	dead	3	3
	alive	4	2
treated	dead	5	15
	alive	8	12

4.2 *Mosaic Plot*



A Greater % Of Those Recieving Treatment Are Alive

4.3 *Pie Chart*



A Greater % Of Those Recieving Treatment Are Alive

5 *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`