M1-06: Simpson’s Paradox

Part of the “Basics of Data Science with Python” Learning Badge

Video Walkthrough:* *https://discovery.cs.illinois.edu/m1-06/

Simpson’s Paradox

Simpson's Paradox refers to a phenomenon in which a trend appears in different groups of data but disappears or reverses when these groups are combined.

In other words, the overall percentages in two groups can be misleading because of a confounder. Once the confounder is controlled for by stratification, the overall effect disappears or is reversed.

Puzzle #1: A 20-year (1974-94) study of 1314 British women compared death rates of smokers to non-smokers and found that 23.9% of the smokers had died compared to 31.4% of the non-smokers.

a) What type of study is this?

b) The data is given below. Could this be evidence that smoking helps you live longer?

	Smokers			Non-Smokers		
Age	**#**	**Deaths**	**Death Rate**	**#**	**Deaths**	**Death Rate**
18-34	179	5	5 / 179 = 2.8%	219	6	6 / 219 = 2.7%
35-64	354	92	92 / 354 = 26.0%	320	59	59 / 320 = 18.4%
65+	49	42	42 / 49 = 85.7%	193	165	165 / 193 = 85.5%
Total	582	139	139 / 582 = 23.9%	732	230	230 / 732 = 31.4%

Summary: Non-smokers had a high overall death rate because a higher percentage of them were old compared to the smokers. The groups were not balanced at the start of the study. If you stratify based on age you see that smokers do have a higher death rate.

Puzzle #2: Can Simpson’s Paradox occur in randomized controlled experiments with large sample sizes? Why or why not?